Combining Information from Multiple Sources in Bayesian Modeling

by

Tracy Anne Schifeling

Department of Statistical Science
Duke University

Date: ____________________________

Approved:

____
Jerome P. Reiter, Advisor

____
Surya Tokdar

____
Fan Li

____
Seth Sanders

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2016

Abstract

(Statistics)

Combining Information from Multiple Sources in Bayesian Modeling

by

Tracy Anne Schifeling

Department of Statistical Science
Duke University

Date: ________________________

Approved:

____
Jerome P. Reiter, Advisor

____
Surya Tokdar

____
Fan Li

____
Seth Sanders

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2016

Abstract

Surveys can collect important data that inform policy decisions and drive social science research. Large government surveys collect information from the U.S. population on a wide range of topics, including demographics, education, employment, and lifestyle. Analysis of survey data presents unique challenges. In particular, one needs to account for missing data, for complex sampling designs, and for measurement error. Conceptually, a survey organization could spend lots of resources getting high-quality responses from a simple random sample, resulting in survey data that are easy to analyze. However, this scenario often is not realistic. To address these practical issues, survey organizations can leverage the information available from other sources of data. For example, in longitudinal studies that suffer from attrition, they can use the information from refreshment samples to correct for potential attrition bias. They can use information from known marginal distributions or survey design to improve inferences. They can use information from gold standard sources to correct for measurement error. This thesis presents novel approaches to combining information from multiple sources that address the three problems described above.

The first method addresses nonignorable unit nonresponse and attrition in a panel survey with a refreshment sample. Panel surveys typically suffer from attrition, which can lead to biased inference when basing analysis only on cases that complete all waves of the panel. Unfortunately, the panel data alone cannot inform the extent of the bias due to attrition, so analysts must make strong and untestable assumptions about the missing data mechanism. Many panel studies also include refreshment samples, which are data collected from a random sample of new individuals during some later wave of the panel. Refreshment samples offer information that can be utilized to correct for biases induced by nonignorable attrition while reducing reliance on strong assumptions about the attrition process. To date, these bias correction methods have not dealt with two key practical issues in panel studies: unit nonresponse in the initial wave of the panel and in the refreshment sample itself. As we illustrate, nonignorable unit nonresponse can significantly compromise the analyst's ability to use the refreshment samples for attrition bias correction. Thus, it is crucial for analysts to assess how sensitive their inferences—corrected for panel attrition—are to different assumptions about the nature of the unit nonresponse. We present an approach that facilitates such sensitivity analyses, both for suspected nonignorable unit nonresponse in the initial wave and in the refreshment sample. We illustrate the approach using simulation studies and an analysis of data from the 2007-2008 Associated Press/Yahoo News election panel study.

The second method incorporates informative prior beliefs about marginal probabilities into Bayesian latent class models for categorical data. The basic idea is to append synthetic observations to the original data such that (i) the empirical distributions of the desired margins match those of the prior beliefs, and (ii) the values of the remaining variables are left missing. The degree of prior uncertainty is controlled by the number of augmented records. Posterior inferences can be obtained via typical MCMC algorithms for latent class models, tailored to deal efficiently with the missing values in the concatenated data. We illustrate the approach using a variety of simulations based on data from the American Community Survey, including an example of how augmented records can be used to fit latent class models to data from stratified samples.

The third method leverages the information from a gold standard survey to model reporting error. Survey data are subject to reporting error when respondents misunderstand the question or accidentally select the wrong response. Sometimes survey respondents knowingly select the wrong response, for example, by reporting a higher level of education than they actually have attained. We present an approach that allows an analyst to model reporting error by incorporating information from a gold standard survey. The analyst can specify various reporting error models and assess how sensitive their conclusions are to different assumptions about the reporting error process. We illustrate the approach using simulations based on data from the 1993 National Survey of College Graduates. We use the method to impute error-corrected educational attainments in the 2010 American Community Survey using the 2010 National Survey of College Graduates as the gold standard survey.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

$\mathbf{X}$	Demographic variables
Y, Z	Outcome variables
n	Size of survey sample
N	Size of population

Abbreviations

MAR	Missing at Random
MCAR	Missing Completely at Random
NMAR	Not Missing at Random
MCMC	Markov Chain Monte Carlo
MI	Multiple imputation
DPMPM	Dirichlet Process Mixture of Product Multinomials

Acknowledgements

This research was supported by grants from the National Science Foundation (NSF-SES-10-61241 and NSF-SES-11-31897). I would like to thank SAMSI (DMS-1127914) and the Computational Methods in the Social Sciences program, which supported my research during the 2013-14 academic year. I also would like to thank the Survey Research Methods Section of the American Statistical Association for funding my travel to the Joint Statistical Meetings of 2014. Lastly, I thank my advisor, Professor Jerry Reiter, for being a great mentor on my path to becoming a researcher.

1

Introduction

Surveys can be very informative for all sorts of questions. Large government surveys, such as the American Community Survey by the Census Bureau, collect information from the U.S. population on a wide range of topics, including transportation, living situation, language spoken at home, and industry of employment. The General Social Survey, a project of the independent research organization NORC at the University of Chicago, collects information on a wide range of social topics, including general happiness, feelings about religion, and attitudes towards gay marriage (Smith et al., 2015). There are many smaller surveys that target a specific subpopulation, such as the National Survey of College Graduates by the National Science Foundation and the National Jewish Population Survey by The Jewish Federations of North America.

Analysis of surveys is complicated in practice. For example, one needs to account for missing data, for complex sampling designs, and for measurement error (Brick and Kalton, 1996). Conceptually, one could spend lots of resources trying to mitigate problems caused by these issues. For example, one could do an expensive follow up operation to handle nonresponse. Or one could spend lots of money to do in-person interviews to get high quality responses. However, these options are not feasible, because of time and resource constraints.

Faced with practical constraints, survey organizations and analysts often leverage the information in other data sources to deal with the problems. As examples relevant to this thesis, in longitudinal studies they use the information in refreshment samples to correct for attrition bias. They use prior information from known margins or survey design to improve inferences. They use information from gold standard sources to correct for measurement error.

In this thesis, I present novel approaches to combining information from multiple sources. Specifically, I develop methods for the three problems described above. In Chapter 2, I consider the analysis of a panel survey that suffers from two forms of missing data: unit nonresponse and attrition. The survey includes refreshment samples, which are data collected from a sample of new individuals at a later wave of the survey. The refreshment sample can suffer from unit nonresponse. It also has missing data by design, because the individuals in the refreshment sample are only given the survey at one particular wave so that their responses at earlier and later waves are unobserved. I present an approach to jointly model the panel survey and the refreshment sample so as to combine information from both sources of data. In particular, the model has certain sensitivity parameters the analyst can adjust to see how sensitive her inferences are to different assumptions about the unit nonresponse in both sources of data.

In Chapter 3, I consider the analysis of a survey with additional information from another source about the marginal distributions of certain variables. For example, suppose we are analyzing a survey with common demographic variables such as gender, race, and age. From another source of data such as the Census, we can obtain a very precise estimate of the proportion of males in the population. To incorporate this information about the marginal distribution of gender, I create a margin of synthetic records such that the empirical distribution of gender matches the external information and the remaining variables are missing. I show how encoding the marginal information as a synthetic margin allows us to model the survey with the augmented records and thus incorporate marginal information. I apply our method to account for stratified sampling design when fitting a latent class model.

In Chapter 4, I consider the problem of accounting for reporting error in surveys. Suppose we have one survey, D_E, that is subject to reporting error in the variable of interest Z. The true value of the variable of interest is denoted Y. The response to Y is missing (by design) for all individuals in D_E because we do not know what their true response is. We have another gold standard survey, D_G, that measures the same variable of interest without error. In other words, for individuals in D_G, we observe the true value Y but the reported value Z is missing by design. I develop an approach to specify the reporting error mechanism that incorporates information from D_G when imputing the variable of interest in D_E.

In the remainder of this chapter, I review some common statistical ideas that recur throughout the thesis.

1.1 Missing data and multiple imputation

Missing data can arise in a number of ways (Brick and Kalton, 1996). It is useful to classify common types of missing data (Little and Rubin, 2002). Some individuals selected to be in the survey sample simply do not respond; this is known as unit nonresponse. Item nonresponse occurs when an individual leaves some questions blank but does respond to others. A longitudinal survey that tracks the same individuals over several time points can suffer from attrition, which refers to respondents dropping out before the final time point (Hogan and Daniels, 2008). Data can also be missing by design, for example if a certain question was not included in the survey or only given to a subset of the sample (Gelman et al., 1998a; Raghunathan and Grizzle, 1995). In fact, we also frame reporting error in a missing data context. We consider the true response to be missing by design, because the survey only provides the reported response.

We closely follow Gelman et al. (2004) in our discussion of missing data terminology. Let Y be the complete survey, with Y_{obs} being the part that is observed, and Y_{mis} being the part that is missing. The missing data could be caused by any of the reasons given above, such as unit nonresponse or attrition.

Let R be an indicator, such that $R = 1$ if the corresponding Y is observed, and $R = 0$ if the corresponding Y is missing. Let ϕ parameterize the response mechanism, $p(R|Y = (Y_{mis}, Y_{obs}), \phi)$. When the probability of responding does not depend on the values that are missing, i.e., $p(R|Y = (Y_{mis}, Y_{obs}), \phi) = p(R|Y_{obs}, \phi)$, the data are said to be missing at random (MAR) (Rubin, 1976). When the missing data mechanism does depend on the missing values, i.e., $p(R|Y = (Y_{mis}, Y_{obs}), \phi)$, the data are said to be not missing not at random (NMAR). This is the most difficult scenario, because the fact that certain values are missing depends on the missing values themselves. Knowledge about the missing data mechanism or external data beyond the original survey, such as a refreshment sample, are needed to make inference about the missing data mechanism when the data are NMAR (Rubin, 1987; Hirano et al., 2001; Imbens and Pizer, 2000).

When the probability of responding does not depend on Y at all, i.e., $p(R|Y = (Y_{mis}, Y_{obs}), \phi) = p(R|\phi)$, the data are said to be missing completely at random (MCAR). For example, suppose that by design, some individuals in a sample where not asked a specific survey question; these values would be missing completely at random. In the refreshment sample in Chapter 2, the individuals are only given the survey at a particular wave and so their responses to other waves are missing by design.

Analyzing survey data is much more straightforward when there is no missing data. For this reason a researcher may wish to create a complete dataset by imputing, or filling in, Y_{mis}. There is uncertainty about the true values of Y_{mis}, and imputing Y_{mis} only once does not capture the distribution of possible Y_{mis}. Rather than imputing Y_{mis} only once, the idea behind multiple imputation is to create M versions of Y_{mis} (Rubin, 1987). The M multiple imputations of Y_{mis} capture the uncertainty about the true value of Y_{mis}.

We use Markov Chain Monte Carlo (MCMC) to fit Bayesian models, and we include a step in the MCMC to draw Y_{mis} from its posterior distribution (Schafer, 1997). This has two purposes. First, we are often interested in obtaining posterior draws of model parameters θ, where $Y \sim f(\theta)$. It might be difficult to sample from $p(\theta|Y_{obs})$ directly. Instead, we can use the method of data augmentation to integrate out Y_{mis}. At each iteration of the MCMC algorithm, we alternate between sampling from the posterior of θ and imputing the missing data by drawing Y_{mis} from its posterior distribution (Hoff, 2009). Each iteration t consists of two main steps.

1. Sample $\theta^{(t)}$ from $p(\theta|Y = (Y_{obs}, Y_{mis}^{(t-1)}))$.

2. Sample $Y_{mis}^{(t)}$ from $p(Y_{mis}|\theta^{(t)}, Y_{obs})$.

The second purpose of imputing Y_{mis} in the MCMC is to create multiple imputations. To obtain independent draws of Y_{mis}, it is recommended to save every kth draw of Y_{mis} where k is large enough that the draws of Y_{mis} are essentially independent (Schafer, 1997). We can save M draws of Y_{mis} in this way, and then analyze the M completed data sets using Rubin's multiple imputation combining rules. Since we often run the MCMC algorithm for a large number of iterations, it is easy to save $M = 50$ or 100 nearly independent draws of Y_{mis}. The standard combining rules are as follows, if we wish to make inference about a parameter Q (Schafer, 1997; Rubin, 1987):

- Compute the point estimate $\hat{Q}_m$ in each completed dataset $m = 1, \ldots, M$. Let $\bar{Q}_M = \frac{1}{M}\sum_{m=1}^{M} \hat{Q}_m$.

- Estimate $\hat{U}_m$, the variance of $\hat{Q}_m$, in each completed dataset. Let $\bar{U}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{U}_m$. This is the within-imputation variance.

- Let $B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{Q}_m - \bar{Q}_M)^2$. This is the between-imputation variance.

- Let $T_M = \left(1 + \frac{1}{M}\right) B_M + \bar{U}_M$. This is the total variance.

- Let $\nu_M = (M-1)\left(1 + \bar{U}_M/((1+1/M)B_M)\right)^2$.

Putting this all together, we can make inference on Q using a t-distribution with mean $\bar{Q}_M$, scale parameter T_M, and ν_M degrees of freedom.

1.2 Data fusion

In this thesis, we often incorporate information from another data source when imputing missing data. We discuss several different patterns of missing data in multiple data sources.

In some situations, we may observe all variables for some individuals. For example, Rubin et al. (1995) and Gelman et al. (2004) analyze three questions in the Slovenian Public Opinion survey of 1990 regarding the respondents' opinion about Slovenian independence. Each of these three questions had answers "yes," "no", or "don't know." A response of "don't know" is considered missing data. In this survey, the majority of the 2074 survey respondents answered "yes" or "no" for all three questions, meaning their responses are fully observed for these questions of interest. The authors impute the missing data using a MAR model.

In other situations, we may not observe all variables for any individual. For example, Gelman et al. (1998a) present a method to combine information from multiple surveys that assumes at least each pair of questions is asked together in one of the surveys (Gelman et al., 1998b). They assume the survey responses follow a multivariate normal distribution, and so the joint distribution of each pair of questions is sufficient to estimate the model. Related work that addresses combining information from multiple surveys when the surveys do not ask all of the same questions include Carrig et al. (2015), Dominici et al. (1997) and Jackson et al. (2009). Raghunathan and Grizzle (1995) use a similar idea to construct a split questionnaire survey design. The questionnaire is split into components, and any individual is given a small number of these components. The survey should be split in such a way that quantities of interest can still be measured even without asking every question of every individual (Raghunathan and Grizzle, 1995).

Data fusion describes the scenario where the variables of interest are not jointly observed for any individuals (D'Orazio et al., 2006; Rubin, 1986; Rassler, 2002; Reiter, 2012; Moriarity and Scheuren, 2001). In this case the data provide no information about the joint distribution of the variables of interest, besides their marginal distributions. A common assumption is that the variables being fused are conditionally independent given the set of variables common to both surveys.

1.3 Bayesian finite population inference

Surveys are often designed using complex sampling designs, such as stratification or clustering. Complex sampling designs often are more convenient that simple random samples, and some types of sampling designs can increase the efficiency of estimates (Lohr, 2010). Given the sampling design, each unit i in the population of size N has some probability π_i of being selected in the sample of size n. The sampling weight of each individual is $w_i = 1/\pi_i$. The sampling weight can be thought of as the number of individuals in the population that individual i represents in the sample. Often the survey weights that are released with survey data are not exactly the same as the initial sampling weights. The final survey weights have often been adjusted to account for nonresponse in the survey or to match known population totals. There are several ways to account for survey weights.

The design-based approach to survey analysis treats the response variable Y as fixed for all individuals in the population, so there is no need to specify a model for Y (Little, 2003). Instead, the sample inclusion indicator variable I, where $I_i = 1$ if individual i is included in the sample and 0 otherwise, is considered a random variable. The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) is an example of a design-based estimator that uses only the survey responses Y_i and sampling weights w_i for individuals $i = 1, \ldots, n$ in the sample to estimate quantities of interest, without specifying any kind of model for Y.

How to use survey weights in Bayesian modeling is not straightforward (Gelman, 2007; Pfeffermann, 1993). A general recommendation is to include all design variables in the model, so as to make the sampling design ignorable (Gelman et al., 2004; Little, 2003). For example, in Chapter 2, we assume the $\boldsymbol{X}$ variables sufficiently describe the sampling design. We estimate the model and then use a Bayesian version of a post-stratified estimator to estimate the population quantity.

Another approach is to incorporate the survey weights into the model. Kunihama et al. (2014) extend a Dirichlet process mixture model to incorporate survey weights, and Si et al. (2015b) develop a Bayesian method to jointly model the survey weights and outcomes.

A different approach is to construct a dataset that can be treated as a simple random sample. One way to do this is to impute the unobserved $N - n$ records in the population given the n observed records in the sample. Schafer (1997) refers to the imputation of non-sampled units as "mass imputation." Zhou et al. (2015), Dong et al. (2014a), and Dong et al. (2014b) extend the weighted finite population Bayesian bootstrap of Cohen (1997) to generate synthetic populations that account for complex sampling design. In the context of synthetic survey data, Reiter (2002) and Raghunathan et al. (2003) create synthetic populations using methods such as Bayesian bootstrap (Rubin, 1987). Other references of imputing the unobserved records that account for survey weights or sampling design include Little and Zheng (2006), Pfeffermann (2011), Schifeling and Reiter (2016), and Lazar et al. (2008).

2

Accounting for Nonignorable Unit Nonresponse and Attrition in Panel Studies with Refreshment Samples

The presentation in this chapter closely follows the work of Schifeling et al. (2015).

2.1 Introduction

Longitudinal or panel surveys, in which the same individuals are interviewed repeatedly at different points in time, have widespread use in political science (Hillygus, 2005), economics (Baltagi and Song, 2006), education (Buckley and Schneider, 2006), and sociology (Western, 2002), among other fields. Inevitably, panel surveys suffer from attrition; that is, units who respond in early waves fail to respond to later waves of the survey (Lynn, 2013). For example, in the most recent multi-wave panel survey of the American National Election Study, 47% of respondents attrited between the January 2008 baseline wave and the June 2010 follow-up wave. As is well known, attrition can result in biased inferences when the propensity to drop out is systematically related to the substantive outcome of interest (Olsen, 2005; Behr et al., 2005;

Hogan and Daniels, 2008). For example, Bartels (1999) showed that differential attrition of respondents in the 1992-1996 American National Election Study panel resulted in an overestimation of political interest in the population.

Many panel surveys also include refreshment samples: cross-sectional, random samples of new respondents given the questionnaire at the same time as a second or subsequent wave of the panel. Refreshment samples offer information that can be leveraged to correct for nonignorable attrition via statistical modeling. Specifically, as described by Hirano et al. (1998, 2001) and Bhattacharya (2008), the analyst can estimate an additive nonignorable (AN) model, which comprises a joint model for the survey variables coupled with a selection model for the attrition process. Recently, Deng et al. (2013) show that the AN model can be extended to panels with more than two waves and one refreshment sample, including scenarios where attriters in one wave return to the sample in a later wave.

To date, applications of the AN model have largely ignored a key complication in panel surveys, namely that the initial panel and the refreshment sample may be subject to unit nonresponse (e.g., sampled individuals refuse to cooperate, cannot be contacted, or are otherwise unable to participate). For example, in their analysis of the Dutch Transportation Panel, Hirano et al. (1998) treated the 2886 households that completed the first interview as the baseline, disregarding that this represented only 47% (2886/6128) of sampled households in that wave. Bhattacharya (2008) uses the Current Population Survey to design illustrative simulations of an AN model without any unit nonresponse in the panel and refreshment sample. Other applications of the AN model that disregard unit nonresponse include those in Nevo (2003) and Das et al. (2011).

Using only respondents in the initial wave and refreshment sample, even if their cross-sectional weights are calibrated to trusted population estimates, implicitly assumes that the unit nonresponse is missing (perhaps completely) at random (MAR).

This implies strong assumptions about the unit nonresponse, e.g., the distributions of survey variables are identical for respondents and nonrespondents within groups defined by cross-tabulations of limited sets of variables. Unfortunately, as we illustrate later, when the unit nonresponse in the initial wave or refreshment sample is not missing at random (NMAR), treating the unit nonresponse as MAR compromises the analyst's ability to use the refreshment samples to correct for attrition bias. Given sharp declines in survey response rates in recent years (e.g., Peytchev, 2013) and heightened concerns about the impact on the sample representativeness (e.g., Groves et al., 2002; Singer, 2006), disregarding unit nonresponse in applications of the AN model seems difficult to justify.

In this chapter, we present an approach to estimating the AN model for two-wave panels with nonignorable unit nonresponse in the initial wave or refreshment sample. The approach facilitates a process for assessing the sensitivity of inferences corrected for panel attrition to various assumptions about the nature of the nonignorable unit nonresponse. Such sensitivity analyses are the best way to handle unit nonresponse, since the data do not offer information about NMAR missing values (Little and Rubin, 2002; Hogan and Daniels, 2008; Molenberghs et al., 2008). The basic idea is to introduce selection models for the unit nonresponse with interpretable sensitivity parameters that can be tuned to reflect various departures from ignorable missing data mechanisms. This approach is similar in spirit to methods used to assess the sensitivity of estimated treatment effects to unmeasured confounding in observational studies (e.g., Rosenbaum, 2010; Schwartz et al., 2012). We present the methodology for binary outcomes and categorical predictors, although similar ideas could be used for other types of variables. We estimate the models using Bayesian methods and Markov Chain Monte Carlo simulation. We also present two extensions: (i) a two-step approximation to the fully Bayesian solution that can facilitate exploration of different unit nonresponse mechanisms, and (ii) an approach for handling nonignorable unit nonresponse when the panel includes more than one wave before the refreshment sample is collected. We apply the methodology to data from the AP-Yahoo News 2008 Election Panel Study (APYN), focusing on measuring campaign interest in the 2008 presidential election.

The remainder of the chapter is organized as follows. In Section 2.2, we offer a review of the AN model assuming no unit nonresponse in the initial wave and the refreshment sample. In Section 2.3 we present methods for assessing sensitivity of results to nonignorable unit nonresponse in both the initial wave and refreshment sample. We present both the fully Bayesian model and two-step approximation. In Section 2.4 we extend the model to a scenario with an intermediate wave between the baseline and refreshment sample where we allow for missing data in the intermediate wave, caused by either unit nonresponse or panel attrition. In Section 2.5, we illustrate the methodology on an analysis of APYN data.

2.2 Review of Additive Nonignorable Model

Consider a two wave panel of n_p individuals with a refreshment sample of n_r new subjects in the second wave. For all $n = n_p + n_r$ subjects, the data include q time-invariant variables $\mathbf{X} = (X_1, \ldots, X_q)$, such as demographic or frame variables. Let Y_1 be a scalar response variable of substantive interest collected in the initial wave of the panel. Let Y_2 represent the corresponding response variable collected in wave 2. Here, we assume that Y_1 and Y_2 are the same variable collected at different waves, although this is not necessary for the AN model. Among the n_p individuals, $n_{cp} < n_p$ provide at least some data in the second wave, and the remaining $n_{ip} = n_p - n_{cp}$ individuals drop out of the panel. Thus, the refreshment sample includes only $(\mathbf{X}, Y_2)$; by design, Y_1 are missing for all the individuals in the refreshment sample. For now, we presume no nonresponse in Y_1 in the panel, in Y_2 in the refreshment sample, and in $\mathbf{X}$ for all cases.

For each individual $i = 1, \ldots, n$, let $R_i = 1$ if individual i would remain in wave 2 if included in wave 1, and let $R_i = 0$ if individual i would drop out of wave 2 if included in wave 1. We note that R_i is fully observed for all individuals in the panel but is missing for the individuals in the refreshment sample, since this latter set is not provided the chance to respond in wave 1.

The AN model requires a joint model for (Y_1, Y_2) and R given $\mathbf{X}$. We write this as

$$\begin{aligned} (Y_1, Y_2) \mid \mathbf{X} &\sim f(\mathbf{X}, \Theta) && (2.1) \\ R \mid Y_1, Y_2, \mathbf{X} &\sim g(\mathbf{X}, Y_1, Y_2, \boldsymbol{\tau}), && (2.2) \end{aligned}$$

where Θ and $\boldsymbol{\tau}$ represent sets of model parameters. In the APYN application, Y_1 and Y_2 are binary and $\mathbf{X}$ is exclusively categorical. In this case, one specification of the AN model is

$$\begin{aligned} Y_{1i} \mid \mathbf{X}_i \sim Bern(\pi_1), &\quad logit(\pi_1) = \alpha_0 + \boldsymbol{\alpha}_X \mathbf{X}_i && (2.3) \\ Y_{2i} \mid Y_{1i}, \mathbf{X}_i \sim Bern(\pi_{2i}), &\quad logit(\pi_{2i}) = \beta_0 + \beta_1 Y_{1i} + \boldsymbol{\beta}_X \mathbf{X}_i && (2.4) \\ R_i \mid Y_{1i}, Y_{2i}, \mathbf{X}_i \sim Bern(\pi_{Ri}), &\quad logit(\pi_{Ri}) = \tau_0 + \tau_1 Y_{1i} + \tau_2 Y_{2i} + \boldsymbol{\tau}_X \mathbf{X}_i. && (2.5) \end{aligned}$$

Here, we use a subscript X to denote a vector of coefficients in front of $\mathbf{X}_i$; for example, $\boldsymbol{\alpha}_X$ represents the vector of coefficients of $\mathbf{X}$ in the model for Y_1. Throughout the chapter, we implicitly assume that the analyst uses dummy coding to represent the levels of each variable in $\mathbf{X}$ and a separate regression coefficient for each dummy variable.

The key assumption in the AN model is that the probability of attrition depends on Y_1 and Y_2 through a function g that is additive in Y_1 and Y_2; that is, no interactions between Y_1 and Y_2 are allowed in (2.2). Additivity is necessary to enable point identification of the model parameters—i.e., the maximum likelihood estimate of the parameters in the model has a unique value—as there is insufficient information to estimate interactions between Y_1 and Y_2 in (2.2). We note that the model also can accommodate interaction terms among subsets of $(\mathbf{X}, Y_1)$ and interaction terms among subsets of $(\mathbf{X}, Y_2)$. For further discussion of the additivity assumption and when it might not hold, including the consequences of incorrectly assuming additivity, see Deng et al. (2013).

As described by Hirano et al. (1998), AN models include MAR and NMAR models as special cases. When $\tau_2 = 0$ and at least one element among $\{\boldsymbol{\tau}_X, \tau_1\}$ does not equal zero, the attrition is MAR. When $\tau_2 \neq 0$, the attrition is NMAR. Hence, the AN model allows the data to decide between MAR and (certain types of) NMAR attrition mechanisms.

To illustrate how the AN assumption enables point identification, consider Y_1 and Y_2 as binary responses without any other variables, i.e., $\mathbf{X}$ is empty. The data then comprise a contingency table with eight cells, $\{Y_1 = y_1, Y_2 = y_2, R = r : y_1, y_2, r \in \{0, 1\}\}$. As described by Deng et al. (2013), the panel alone yields six constraints on these eight cells, namely (i) the four values of $P(Y_1 = y_1, Y_2 = y_2, R = 1)$ for each combination of $y_1, y_2 \in \{0, 1\}$, (ii) the equation $P(Y_1 = 1, Y_2 = 0, R = 0) + P(Y_1 = 1, Y_2 = 1, R = 0) = P(Y_1 = 1, R = 0)$, and (iii) the requirement that the total probability in the eight cells sums to one. The refreshment sample adds an additional constraint, which can be expressed via the equation $P(Y_2 = 1) = P(Y_2 = 1, R = 0) + P(Y_2 = 1, R = 1)$. Thus, the combined panel and refreshment data can identify models with six parameters plus the sum to one constraint. Excluding coefficients in front of $\mathbf{X}$, this is exactly the number of parameters in (3) – (5); hence, the AN model is point-identified.

We note that AN models can be constructed for outcomes that are not binary. For example, Hirano et al. (1998) present an AN model for normally distributed outcomes, and Bhattacharya (2008) presents an approach to estimating conditional expectations under an AN assumption. In this chapter, we consider AN models based on logistic regressions as in (3) – (5); hence, in addition to the no-interactions assumption, we implicitly assume that logistic regressions are reasonable statistical models for the outcomes of interest.

2.3 Unit Nonresponse in the Baseline or Refreshment Sample

We now describe how to adapt the AN model when there is nonignorable unit nonresponse in the initial wave and in the refreshment sample. Our adaptations facilitate investigations of the sensitivity of inferences to the nonignorable nonresponse. Here, we assume that units that do not respond in the initial wave do not have the opportunity to respond in future waves, as is common in panel surveys. We describe the approach using the models in (3) – (5), but the ideas apply more generally.

For all n_p units in the original panel, let $W_{1i} = 1$ if unit i responds to the initial wave of the panel, and $W_{1i} = 0$ otherwise. For all n_r units in the refreshment sample, let $W_{2i} = 1$ if unit i responds in the refreshment sample, and let $W_{2i} = 0$ otherwise. Here, W_{1i} and W_{2i} describe missingness in Y_{1i} and Y_{2i}, respectively. To simplify explanations, for now we assume $\mathbf{X}_i$ is known for all n units. In practice, this could arise when $\mathbf{X}_i$ comprises sampling frame or administrative variables, or in surveys that have previously collected demographic information, such as internet panels. We discuss relaxing this assumption about $\mathbf{X}_i$ in Section 2.5. Figure 2.1 displays the pattern of observed and missing data across the two waves and refreshment sample.

Figure 2.1 reveals several key features about W_1 and W_2. First, W_1 and W_2 are never observed jointly. Thus, when modeling, we cannot identify any parameter reflecting an association between W_1 and W_2 given the other variables. Second, since we do not observe R and W_2 jointly, we cannot identify any parameter reflecting an association between R and W_2 given the other variables. Third, since we do not observe R when $W_1 = 0$, we cannot identify any parameter reflecting an association between R and W_1 given the other variables.

	$\mathbf{X}$	Y_1	W_1	Y_2	R	W_2
Panel	✓	✓	1	✓	1	?
				?	0	?
	✓	?	0	?	?	?
Refreshment sample	✓	?	?	✓	?	1
	✓	?	?	?	?	0

FIGURE 2.1: Two-wave panel with attrition and unit nonresponse in the baseline and refreshment sample.

Based on these constraints, we add the following selection models to (2.3), (2.4), and (2.5).

$$W_{1i}|Y_{1i}, \mathbf{X}_i \sim Bern(\pi_{W_{1i}}), \quad logit(\pi_{W_{1i}}) = \gamma_0 + \gamma_1 Y_{1i} + \boldsymbol{\gamma}_X \mathbf{X}_i \tag{2.6}$$

$$W_{2i}|Y_{2i}, \mathbf{X}_i \sim Bern(\pi_{W_{2i}}), \quad logit(\pi_{W_{2i}}) = \rho_0 + \rho_1 Y_{2i} + \boldsymbol{\rho}_X \mathbf{X}_i. \tag{2.7}$$

We let γ_1 and ρ_1 be sensitivity parameters set by the analyst, and $(\gamma_0, \boldsymbol{\gamma}_X)$ and $(\rho_0, \boldsymbol{\rho}_X)$ be estimated from the data.

For illustration, we return to the scenario with Y_1 and Y_2 binary and $\mathbf{X}$ empty. In this case, the combined panel and refreshment sample essentially represent a contingency table with $2^5 = 32$ cells. We do not observe counts in any of the cells directly because each unit has at least one variable missing, either by design or because the unit did not respond. For example, even for the panel units who respond in both wave one and wave two, we do not know whether they would have responded had they been in the refreshment sample, i.e., their values of W_2 are unobserved. As a consequence, the ten parameters in (2.3) through (2.7) are not point-identified, as we now explain.

The complete-cases in the panel data provide four constraints, for $y_1, y_2 \in \{0, 1\}$, namely

$$P(Y_1 = y_1, Y_2 = y_2, W_1 = 1, R = 1)$$
$$= \sum_{w_2 \in \{0,1\}} P(Y_1 = y_1, Y_2 = y_2, W_1 = 1, R = 1, W_2 = w_2). \tag{2.8}$$

The attriters from the cases with $W_1 = 1$ offer two constraints, for $y_1 \in \{0, 1\}$, namely

$$\begin{aligned} P(Y_1 = y_1, W_1 = 1, R = 0) & \\ &= \sum_{y_2, w_2} P(Y_1 = y_1, Y_2 = y_2, W_1 = 1, R = 0, W_2 = w_2). \quad (2.9) \end{aligned}$$

The respondents in the refreshment sample data provide another two constraints, for $y_2 \in \{0, 1\}$, namely

$$P(Y_2 = y_2, W_2 = 1) = \sum_{y_1, w_1, r} P(Y_1 = y_1, Y_2 = y_2, W_1 = w_1, R = r, W_2 = 1). \quad (2.10)$$

Thus, the data offer only eight constraints to estimate ten parameters (ignoring the sum-to-one constraint, which is present in the data and the complete model).

To enable identification of the parameters in the complete model, we have to add two constraints to (2.3) through (2.7). We do so by fixing γ_1 and ρ_1 at user-specified constants; that is, we make them sensitivity parameters. These can be readily interpreted in terms of odds of response, with values far from zero indicating nonignorable nonresponse. For example, setting $\gamma_1 = 0$ implies that the nonresponse in Y_1 in the panel is missing at random (MAR) and can be ignored, whereas setting $\gamma_1 = .5$ implies that individuals with $Y_1 = 1$ have $\exp(.5) \approx 1.65$ times the odds of responding in wave 1 as individuals with $Y_1 = 0$. Similarly, setting $\rho_1 = 0$ implies that the nonresponse in the refreshment sample is MAR, whereas setting $\rho_1 = -.25$ implies that individuals with $Y_2 = 1$ have $\exp(-.25) \approx .77$ times the odds of responding in the refreshment sample as individuals with $Y_2 = 0$. Of course, it is not possible for the analyst to know γ_1 or ρ_1. However, as we illustrate in Section 2.5.2, the analyst can insert a range of values of each parameter into (2.3) through (2.7) to assess the sensitivity of analyses to assumptions about the nonignorable nonresponse.

With set values of the sensitivity parameters, the model can be estimated using Gibbs sampling. Given a current draw of the completed-data, we sample the parameters using Metropolis steps. Given a current draw of the parameters, we draw values of the missing data according to the logistic regressions in (2.3) through (2.7) as follows. For ease of notation, we suppress parameters from the conditional densities.

- Cases with $(W_1 = 1, R = 0)$: Sample the missing (Y_2, W_2) from

  $$p(Y_2, W_2 \mid \mathbf{X}, Y_1, R = 0)$$
  $$= p(Y_2 \mid \mathbf{X}, Y_1, R = 0) p(W_2 \mid \mathbf{X}, Y_1, Y_2, R = 0) \quad (2.11)$$

  where
  $$p(Y_2 \mid \mathbf{X}, Y_1, R = 0) \propto p(R = 0 \mid \mathbf{X}, Y_1, Y_2) p(Y_2 \mid \mathbf{X}, Y_1) \quad (2.12)$$

  and $p(W_2 \mid \mathbf{X}, Y_1, Y_2, R = 0)$ is given by (2.7).

- Cases with $W_1 = 0$: Sample the missing (Y_1, Y_2, W_2, R) from

  $$p(Y_1, Y_2, W_2, R \mid \mathbf{X}, W_1 = 0)$$
  $$= p(Y_1 \mid \mathbf{X}, W_1 = 0) p(Y_2 \mid \mathbf{X}, Y_1) p(R \mid \mathbf{X}, Y_1, Y_2) p(W_2 \mid \mathbf{X}, Y_2) \quad (2.13)$$

  where
  $$p(Y_1 \mid \mathbf{X}, W_1 = 0) \propto p(W_1 = 0 \mid \mathbf{X}, Y_1) p(Y_1 \mid \mathbf{X}) \quad (2.14)$$

  and the remaining densities from (2.4), (2.5), and (2.7).

- Cases with $W_2 = 1$: Sample the missing (Y_1, W_1, R) from

  $$p(Y_1, W_1, R \mid \mathbf{X}, Y_2, W_2 = 1)$$
  $$= p(Y_1 \mid \mathbf{X}, Y_2) p(W_1 \mid \mathbf{X}, Y_1) p(R \mid \mathbf{X}, Y_1, Y_2) \quad (2.15)$$

  where
  $$p(Y_1 \mid \mathbf{X}, Y_2, W_2 = 1) \propto p(Y_2 \mid \mathbf{X}, Y_1) p(Y_1 \mid \mathbf{X}) \quad (2.16)$$

  and the remaining densities from (2.5) and (2.6).

- Cases with $W_2 = 0$: Sample the missing (Y_1, Y_2, W_1, R) from

$$\begin{aligned} p(Y_1, Y_2, W_1, R \mid \mathbf{X}, W_2 = 0) = p(Y_1 \mid \mathbf{X}, W_2 = 0) p(W_1 \mid \mathbf{X}, Y_1) \\ \times\, p(Y_2 \mid \mathbf{X}, Y_1, W_2 = 0) p(R \mid \mathbf{X}, Y_1, Y_2) \end{aligned} \tag{2.17}$$

where

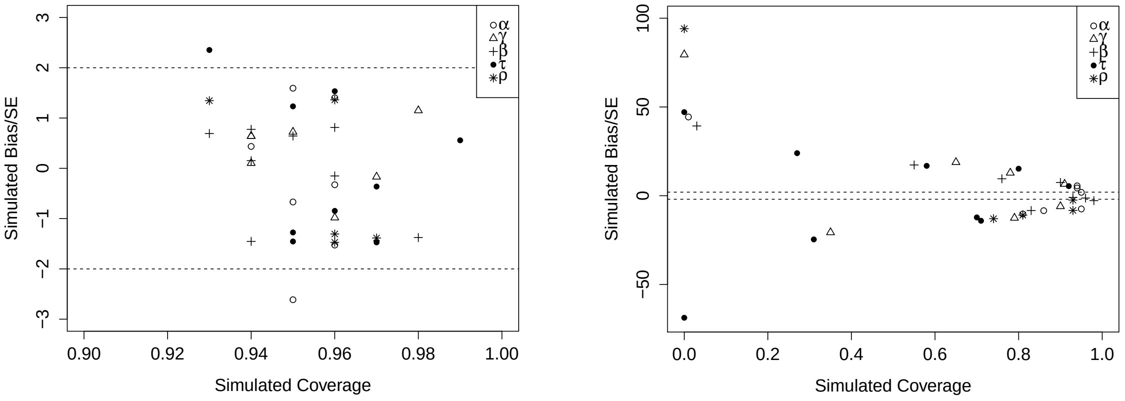
$$p(Y_2 \mid \mathbf{X}, Y_1, W_2 = 0) \propto p(W_2 = 0 \mid \mathbf{X}, Y_1, Y_2) p(Y_2 \mid \mathbf{X}, Y_1), \tag{2.18}$$

$$\begin{aligned} p(Y_1 | \mathbf{X}, W_2 = 0) \propto \sum_{y_2} \Big[p(W_2 = 0 | \mathbf{X}, Y_2 = y_2) \\ \times\, p(Y_2 = y_2 | \mathbf{X}, Y_1) p(Y_1 | \mathbf{X}) \Big] \end{aligned} \tag{2.19}$$

and all densities are from (2.3) through (2.7).

We now demonstrate how to account for nonignorable nonresponse in the baseline and refreshment sample, as well as nonignorable attrition. To do so, we design simulations in which we use the true values of γ_1 and ρ_1. In a sensitivity analysis in a genuine setting, the analyst would not know the true values of the sensitivity parameters and would examine a range of plausible values; we illustrate this in Section 2.5.2. Here, our purpose is to show that the approach can result in accurate parameter estimates when true values are used, thus demonstrating that the approach offers a meaningful sensitivity analysis.

We simulate 100 datasets from our model in (2.3) through (2.7), and estimate the parameters for each dataset under two conditions: (i) with the sensitivity parameters both set to their true values, and (ii) with the sensitivity parameters both set to 0. For each simulated dataset, the panel has 10,000 units and the refreshment sample has 5,000 units. We include eight binary variables, six of which comprise $\mathbf{X}$. We set the values of the logistic regression coefficients to generate significant associations among the survey variables and substantial, but potentially realistic, nonresponse rates. Specifically, the parameters in the model in (2.3) through (2.7) are all set to values between -0.5 and 1. Both sensitivity parameters γ_1 and ρ_1 are set to 1.



(a) Two wave simulation results with sensitivity parameters set to truth.

(b) Two wave simulation results with sensitivity parameters set to zero.

FIGURE 2.2: The left panel displays simulated coverage versus simulated Bias/SE with sensitivity parameters set to truth for model in (3)–(7). The right panel displays results with sensitivity parameters set to zero.

These parameter settings lead to roughly 50% of the panel completing both waves, 25% of the panel dropping out at wave two, and 25% of the panel not responding in wave one. The refreshment sample has roughly a 60% response rate. We use independent Cauchy priors with a scale parameter of 10 for the intercept terms and a scale parameter of 2.5 for all other coefficients, following the advice of Gelman et al. (2008).

As evident in Figure 2.2a, when the sensitivity parameters are set to their true values, for all parameters the simulated coverage rates are between 0.93 and 0.99. On the other hand, as evident in Figure 2.2b, when the sensitivity parameters are set wrongly to zero (which corresponds to MAR nonresponse), 25 parameters have simulated coverage less than 0.9. The five parameters in the top left of Figure 2.2b are the five intercepts, and the parameter in the bottom left is the coefficient of Y_2 in the model for R.

We also ran simulations for both cases where one sensitivity parameter is set to the truth and the other is set to zero. When we correctly account for the unit nonresponse in the baseline but incorrectly assume the refreshment sample unit nonresponse is MAR, some parameters in models for Y_2, R and W_2 are far from their true values. When we correctly account for the unit nonresponse in the refreshment sample but incorrectly assume the baseline unit nonresponse is MAR, some parameters in the models for Y_1 and W_1 are far from from their true values. Thus, when unit nonresponse in both the baseline and refreshment sample is nonignorable, it is important to assess sensitivity to assumptions about both sources of missing data.

We also investigated the performance of the sensitivity analysis procedure when the underlying models are misspecified. Specifically, we generated data from robit regression models (Liu, 2004; Gelman et al., 2004; Gelman and Hill, 2007) with one degree of freedom using the same parameter values as before, but fit the sequential logistic regression models to the data. When we set the sensitivity parameters to the values used previously, inferences continue to be more reliable than when setting the sensitivity parameters to zero. See Section 2.7 for details and results.

Although the selection models in (2.6) and (2.7) are flexible, some analysts may prefer to characterize the missing data in the baseline or refreshment sample with alternate missing data mechanisms, for example, pattern mixture models. In such cases, the analyst can implement a two-step approach based on multiple imputation (Rubin, 1987). First, the analyst fills in the missing data caused by unit nonresponse in the baseline or refreshment sample, creating M completed datasets. After the baseline and refreshment sample are completed, we have only panel attrition, which can be handled with an AN model. Inferences can proceed via the usual multiple imputation combining rules (Rubin, 1987) or, for Bayesian analyses based on (2.3) through (2.5), by mixing the draws of parameters from the M chains (Zhou and Reiter, 2010). We ran simulation studies of this approach basing completions of the nonresponse on the selection models in (2.6) and (2.7); results were similar to those seen in Figure 2.2a and Figure 2.2b.

While we focus on binary outcomes, similar approaches could be used for sensitivity analyses with multinomial or continuous outcomes. For example, if Y_1 is continuous, the analyst could specify a sensitivity parameter for Y_1 (or some function of it) and interpret it using the usual change in log-odds ratio. If Y_1 is a categorical variable with $d > 2$ levels, we could replace γ_1 with $(\gamma_{1,1}, \ldots, \gamma_{1,d-1})$, potentially allowing each value of Y_1 to have a different effect on the probability of responding. With large d, the analyst most likely would need to make simplifying assumptions about the $(\gamma_{1,1}, \ldots, \gamma_{1,d-1})$ to use the approach in practice; for example, set $\gamma_{1,j} = 0$ for some set of j and $\gamma_{1,j} = \gamma_{1,j'}$ for the complementary set.

Finally, while we prefer to examine inferences at particular values of the sensitivity parameters, it is also possible to specify informative prior distributions on these parameters. These could be based on expert opinion or previous experience about the nature of the unit nonresponse. We illustrate this approach in the analysis of the APYN data in Section 2.5.3.

2.4 Intermediate Waves

In some panel surveys, multiple waves may take place between the collection of the baseline data and refreshment sample. These intermediate waves can be subject to panel attrition. In this section, we show that the ideas of Section 2.3 can be extended to correct for nonignorable attrition in intermediate waves as well. We use a setting with three waves including a refreshment sample at the third wave only. We consider two variations: (i) monotone dropout, when those units that do not respond in wave 2 do not have the opportunity to respond in wave 3, and (ii) intermittent dropout, when units that do not respond in wave 2 are potentially able to respond in wave 3. This latter scenario often arises in online panel surveys, like the APYN election poll, in which all wave 1 respondents are invited to participate in subsequent surveys, even if they failed to participate in a previous wave. For simplicity, as in the original AN model, we assume there is no unit nonresponse in the baseline nor in the refreshment sample. Adding selection models like those in Section 2.3 does not present additional complications.

For both scenarios, we use the notation of Section 2.3 with slight modification. For $i = 1, \ldots, n$, let $W_{2i} = 1$ if individual i provides a value of Y_2 if included in wave 1. Note that W_2 is missing for cases in the refreshment sample. For $i = 1, \ldots, n$, let Y_{3i} be the response of unit i at wave 3; let $R_i = 1$ if individual i would provide a

					Monotone		Intermittent	
	$\mathbf{X}$	Y_1	Y_2	W_2	Y_3	R	Y_3	R
Panel	✓	✓	✓	1	✓	1	✓	1
					?	0	?	0
			?	0	?	?	✓	1
							?	0
Refreshment sample:	✓	?	?	?	✓	?	✓	?

Figure 2.3: Two examples of a three wave study with a refreshment sample in the third wave and dropout in the intermediate wave. With monotone dropout, panel units that fail to respond in wave 2 ($W_2 = 0$) do not have the opportunity to continue in wave 3. With intermittent dropout, panel units that fail to respond in wave 2 can participate in wave 3. In either scenario, the patterns of observed and missing data are the same for the baseline wave and refreshment samples.

value in wave 3 if included in wave 2; and, let $R_i = 0$ if individual i would drop out of wave 3 if included in wave 2. Here, R_i is not observed for any units that are not followed up from wave 2, which includes cases in the refreshment sample and with $W_2 = 0$ for monotone dropout. The two variations are displayed in Figure 2.3.

2.4.1 Monotone Dropout

Comparing Figure 2.3 with Figure 2.1, it is apparent that the case with monotone dropout has similar structure as the case for a two-wave panel with nonresponse in the baseline and complete data in the refreshment sample. In particular, consider (Y_1, Y_2) combined as variables in a hypothetical "wave 1" and Y_3 as a hypothetical "wave 2." In this case, only some of the variables in the "wave 1" are subject to nonignorable nonresponse. Taking advantage of this mapping, we can write a model for the monotone dropout case using the general format,

$$(Y_1, Y_2, Y_3) \mid \mathbf{X} \sim f(\mathbf{X}, \Theta) \tag{2.20}$$

$$R \mid Y_1, Y_2, Y_3, \mathbf{X} \sim g(\mathbf{X}, Y_1, Y_2, Y_3, \boldsymbol{\rho}) \tag{2.21}$$

$$W_2 \mid Y_1, Y_2, \mathbf{X} \sim h(\mathbf{X}, Y_1, Y_2, \boldsymbol{\gamma}), \tag{2.22}$$

where Θ, $\boldsymbol{\rho}$, and $\boldsymbol{\gamma}$ represent sets of model parameters. As in the AN model, to enable model identification we exclude interactions between Y_3 and (Y_1, Y_2) when specifying (2.21).

In the case of binary outcomes, we can write this as a sequence of logistic regressions,

$$Y_{1i} \mid \mathbf{X}_i \sim Bern(\pi_1), \quad logit(\pi_1) = \alpha_0 + \boldsymbol{\alpha}_X \mathbf{X}_i \tag{2.23}$$

$$Y_{2i}|Y_{1i}, \mathbf{X}_i \sim Bern(\pi_{2i}), \quad logit(\pi_{2i}) = \beta_0 + \beta_1 Y_{1i} + \boldsymbol{\beta}_X \mathbf{X}_i \tag{2.24}$$

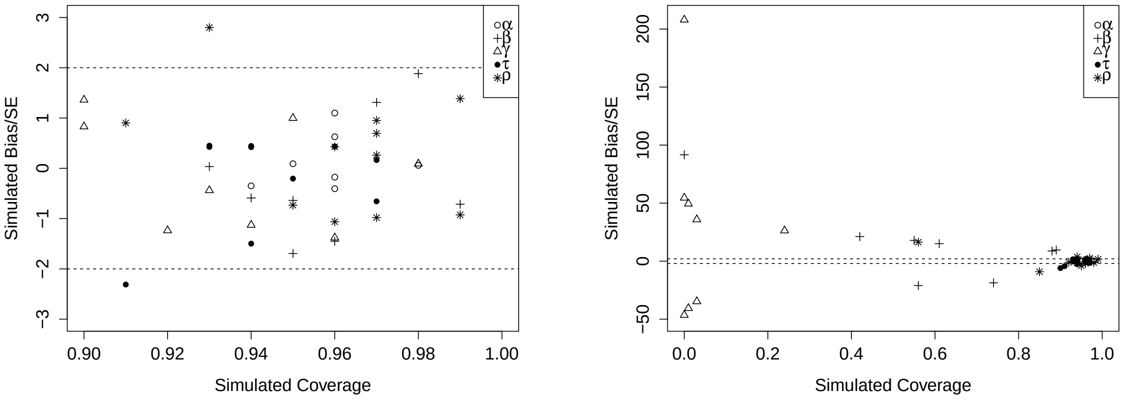
$$\begin{aligned} Y_{3i}|Y_{1i}, Y_{2i}, \mathbf{X}_i \sim Bern(\pi_{3i}), \quad logit(\pi_{3i}) = \Big(&\tau_0 + \tau_1 Y_{1i} + \tau_2 Y_{2i} \\ &+ \tau_3 Y_{1i} Y_{2i} + \boldsymbol{\tau}_X \mathbf{X}_i\Big) \end{aligned} \tag{2.25}$$

$$\begin{aligned} R_i|Y_{1i}, Y_{2i}, Y_{3i}, \mathbf{X}_i \sim Bern(\pi_{R_i}), \quad logit(\pi_{R_i}) = \Big(&\rho_0 + \rho_1 Y_{1i} + \rho_2 Y_{2i} + \rho_3 Y_{3i} \\ &+ \rho_4 Y_{1i} Y_{2i} + \boldsymbol{\rho}_X \mathbf{X}_i\Big) \end{aligned} \tag{2.26}$$

$$W_{2i}|Y_{1i}, Y_{2i}, \mathbf{X}_i \sim Bern(\pi_{W_{2i}}), \quad logit(\pi_{W_{2i}}) = \gamma_0 + \gamma_1 Y_{1i} + \gamma_2 Y_{2i} + \boldsymbol{\gamma}_X \mathbf{X}_i \tag{2.27}$$

Here, γ_2 is a sensitivity parameter that is set to reflect nonignorable response at wave 2. We can include interactions between Y_1 and Y_2 in (2.25) and (2.26) since we have cases with all these variables observed. We note that Hogan and Daniels (2008) discuss a similar model without refreshment samples, treating both γ_2 and ρ_3 as sensitivity parameters. As in the AN model in Section 2.2, the refreshment sample provides enough information to identify ρ_3. Deng et al. (2013) show that with an additional refreshment sample at wave two, γ_2 would be identifiable from the data as well.

We now illustrate via simulations that the model can adjust for nonignorable monotone dropout in an intermediate wave and, hence, can offer valid sensitivity analyses. We simulate 100 datasets from the model in (2.23) through (2.27) and use a Metropolis-within-Gibbs sampler to estimate the parameters for each dataset under two conditions: (i) with the sensitivity parameter set to its true value, and (ii) with the sensitivity parameter set to 0, meaning we assume Y_2 is missing at random. For each simulated dataset, the panel has 10,000 units and the refreshment sample has 5,000 units. We include six binary covariates. We again set the values of



(a) Monotone dropout simulation results with sensitivity parameter set to truth.

(b) Monotone dropout simulation results with sensitivity parameter set to zero.

Figure 2.4: The left panel displays simulated coverage versus simulated Bias/SE with sensitivity parameter set to truth for monotone dropout model in (23)–(27). The right panel displays results with sensitivity parameter set to zero.

the coefficients to produce reasonable nonresponse rates and associations among the variables. The parameters in the the model in (2.23) through (2.27) are all set to values between -0.6 and 0.6, except for the sensitivity parameter γ_2 which is set to 3. These parameter settings lead to roughly 50% of the panel completing all three waves, 25% of the panel dropping out at wave 2, and 25% of the panel dropping out in wave 3. Again we use independent Cauchy priors with a scale parameter of 10 for the intercept terms and a scale parameter of 2.5 for all other coefficients (Gelman et al., 2008). We should note that one of the 100 simulated trials produced obviously invalid results, we suspect due to lack of convergence of the MCMC. We replaced this rogue trial with another run.

As evident in Figure 2.4a, when we set $\gamma_2 = 3$, the simulated coverage rate for all parameters is at least 90%, When we instead set $\gamma_2 = 0$, incorrectly assuming Y_2 is missing at random, many parameters have low simulated coverage rates, especially the parameters in the models associated with wave two. The parameter in the top left of Figure 2.4b with the largest bias is the γ intercept.

2.4.2 Intermittent Dropout

Under intermittent dropout, units that do not respond in wave two still are given the opportunity to respond in wave three. As a result, we now observe $W_2 = 0$ jointly with both Y_3 and R. This information offers four additional constraints, so that we can add four parameters to (2.23) through (2.27). In particular, we add terms for W_2 and Y_1W_2 in the logit equations in (2.25) and (2.26). We have

$$\begin{aligned} logit(\pi_{3i}) = \Big(&\tau_0 + \tau_1 Y_{1i} + \tau_2 Y_{2i} + \tau_3 Y_{1i}Y_{2i} \\ &+ \tau_4 W_{2i} + \tau_5 W_{2i}Y_{1i} + \boldsymbol{\tau}_X \mathbf{X}_i\Big) \end{aligned} \tag{2.28}$$

$$\begin{aligned} logit(\pi_{R_i}) = \Big(&\rho_0 + \rho_1 Y_{1i} + \rho_2 Y_{2i} + \rho_3 Y_{3i} + \rho_4 Y_{1i}Y_{2i} \\ &+ \rho_5 W_{2i} + \rho_6 W_{2i}Y_{1i} + \boldsymbol{\rho}_X \mathbf{X}_i\Big). \end{aligned} \tag{2.29}$$

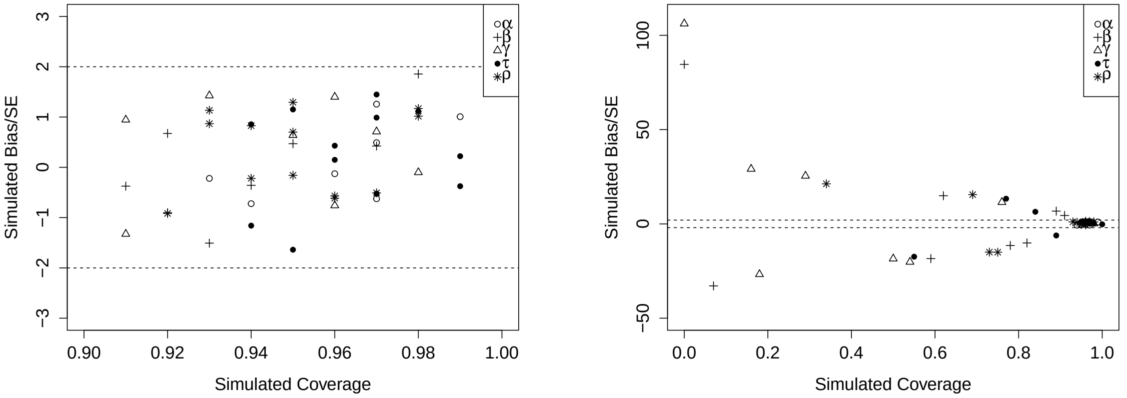
Here, γ_2 remains a sensitivity parameter that is set by the analyst.

To illustrate, we run simulations with the same dimension and sample size as in the monotone dropout simulation study. The parameters in the model in (2.23), (2.24), (2.28), (2.29), and (2.27) are all between -0.4 and 0.6, except for the sensitivity parameter γ_2 which is set to 1.5. These parameter settings lead to about 45% of the panel completing all three waves, about 27% not responding in wave 3 only, 17% not responding in wave 2 only, and 11% not responding in wave 2 or 3.

As evident in Figure 2.5a and Figure 2.5b, when $\gamma_2 = 1.5$, all the parameters have near nominal simulated coverage rates and unremarkable simulated bias. When γ_2 is wrongly set to zero, many parameters have coverage rate far less than 90%. The two parameters with largest bias are the intercepts β_0 and γ_0.

2.5 Dealing with Attrition and Nonresponse in an Analysis of the APYN

The APYN is an eleven-wave panel survey with three refreshment samples intended to measure attitudes about the 2008 presidential election. The panel was sampled from the probability-based KnowledgePanel Internet panel, which recruits panel



(a) Intermittent dropout simulation results with sensitivity parameter set to truth.

(b) Intermittent dropout simulation results with sensitivity parameter set to zero.

FIGURE 2.5: The left panel displays simulated coverage versus simulated Bias/SE with sensitivity parameter set to truth in intermittent dropout model in (23), (24), (28), (29), and (27). The right panel displays results with sensitivity parameter set to zero.

members via a probability-based sampling method using known published sampling frames that cover 96% of the U.S. population. Sampled non-internet households are provided a laptop computer or MSN TV unit and free internet service. The study was a collaboration between the AP and Yahoo Inc., with support from Knowledge Networks (KN).

We analyze two of the eleven waves, specifically wave 1 and wave 9. Wave 1 was fielded on November 2, 2007 to $n_p = 3548$ panelists at least 18 years old, out of whom only 2730 completed the interview (i.e., $\sum W_{1i} = 2730$). Wave 9, the last wave to include a refreshment sample, was fielded in October 2008. Individuals who failed to participate in the first wave of the survey (i.e., with $W_{1i} = 0$) were not subsequently included in the follow up waves. However, individuals who completed the first wave (i.e., $W_{1i} = 1$) were invited to participate in all subsequent waves, even if they skipped one or more of the follow-up waves. Of the 2730 wave 1 respondents, only $n_{cp} = 1715$ remain in the panel by wave 9. The refreshment sample at wave 9 was fielded to $n_r = 1085$ individuals, of whom 461 responded. For the remainder of the analysis, we refer to the November 2007 wave as "wave one" and the October 2008 wave as "wave two."

Following Deng et al. (2013), we analyze campaign interest, which is a strong predictor of democratic attitudes and behaviors (Prior, 2010), is used in identifying likely voters in pre-election polls (Traugott and Tucker, 1984), and has been shown to be susceptible to panel attrition bias (Bartels, 1999). Our outcome of interest is based on the survey question, "How much thought, if any, have you given to candidates who may be running for president in 2008?" Following common usage of this measure (e.g., Traugott and Tucker, 1984), we dichotomize the response into those who respond "A lot" and all other responses, so that $Y_{ti} = 1$ if unit i at wave $t \in \{1, 2\}$ responds "a lot," and $Y_{ti} = 0$ otherwise. In wave 1, 29.8% of the 2730 respondents answer $Y_{1i} = 1$. The percentage with $Y_{2i} = 1$ increases dramatically to 65.0% among the 1715 complete cases at wave two. In the refreshment sample, 72.2% of the 461 respondents answer $Y_{2i} = 1$. We predict campaign interest from age (four categories), education (college degree or not), gender (male or not), and race (black or not); see Table 2.1 for summaries of these variables. We note that Deng et al. (2013) ignored unit nonresponse in the refreshment samples and original panel entirely, effectively acting as if this nonresponse was MCAR.

2.5.1 Missing Demographic Characteristics

In the model and simulation study in Section 2.3, we assume that all $\mathbf{X}$ values are known, even for units in the panel and refreshment sample that do not participate. In the APYN data we analyze, however, we do not have the demographic variables for the 818 panel units with $W_{1i} = 0$ nor for the 624 refreshment sample units with $W_{2i} = 0$. We use the following strategy to create a dataset with complete values of all $\mathbf{X}$.

Let $(\mathbf{x}_1, \ldots, \mathbf{x}_{32})$ represent each of the $4 \times 2 \times 2 \times 2 = 32$ combinations of $\mathbf{X}$ defined by cross-tabulating the four demographic characteristics. We compute the weighted sample proportions from the 2012 Current Population Survey (CPS), which we call $\bar{x}_k^{CPS}$, where $k = 1, \ldots, 32$ indexes the combinations. Let $\bar{x}_k^{PAN} = \sum_{i=1}^{n_p} I(\mathbf{X}_i =$

$\mathbf{x}_k)/n_p$, where $I(\cdot) = 1$ when the condition inside the parentheses is true and $I(\cdot) = 0$ otherwise. These are the completed-data proportions in the panel, computed after imputation of missing $\mathbf{X}$. Similarly, let the completed-data proportions in the refreshment sample be $\bar{x}_k^{REF} = \sum_{i=1}^{n_r} I(\mathbf{X}_i = \mathbf{x}_k)/n_r$, where $k = 1, \ldots, 32$. We impute missing $\mathbf{X}$ values in the panel and refreshment sample so that $\bar{x}_k^{PAN}$ and $\bar{x}_k^{REF}$ closely match $\bar{x}_k^{CPS}$ for all k. We do so because the CPS and APYN both target the population of U.S. adults. Let $n_{pm} = \sum_{i=1}^{n_p}(1 - W_{1i})$ be the number of cases in the panel with missing $\mathbf{X}$ values. Similarly, let $n_{rm} = \sum_{i=1}^{n_r}(1 - W_{2i})$. Using the n_p panel cases, for $k = 1, \ldots, 32$ let $\tilde{x}_k^{PAN} = \sum_{i=1}^{n_p} W_{1i} I(\mathbf{X}_i = \mathbf{x}_k)/n_p$. Define the corresponding quantities for the refreshment sample as $\tilde{x}_k^{REF} = \sum_{i=1}^{n_r} W_{2i} I(\mathbf{X}_i = \mathbf{x}_k)/n_r$. We use the following imputation algorithm.

1. Set a counter $t = 0$.

2. Find k such that $\bar{x}_k^{CPS} - \tilde{x}_k^{PAN}$ is maximized.

3. Among cases with $W_{1i} = 0$ yet to have $\mathbf{X}$ imputed, determine how many additional cases to set the missing $\mathbf{X}_i = \mathbf{x}_k$ so that $|\bar{x}_k^{CPS} - \bar{x}_k^{PAN}|$ is minimized. Call this n_k. Let $t = t + n_k$.

4. Among cases with $W_{1i} = 0$ yet to have $\mathbf{X}$ imputed, set n_k cases values of $\mathbf{X}$ equal to $\mathbf{x}_k$. When $t > n_{pm}$, only impute $\mathbf{X}$ values for the remaining $(n_{pm} - t + n_k)$ cases.

5. Go back to step 2 until all cases with missing $\mathbf{X}$ in the panel have been imputed.

6. Repeat from step 1 for the refreshment sample, replacing (n_p, n_{pm}, W_{1i}) with (n_r, n_{rm}, W_{2i}) and $(\bar{x}_k^{PAN}, \tilde{x}_k^{PAN})$ with $(\bar{x}_k^{REF}, \tilde{x}_k^{REF})$.

Table 2.1 displays the marginal probabilities for each demographic variable before and after the imputation. The imputation scheme generates a completed-data population that mimics the CPS marginal frequencies. When no data source is representative

Table 2.1: Marginal probabilities of the demographic variables in the panel and refreshment sample before and after imputing the missing covariates. When imputing the missing demographic variables, we did so to match joint probabilities available in the 2012 Current Population Survey.

		Panel		Refreshment		
Variable	Definition	Before	After	Before	After	CPS
AGE0	= 1 for age 18-29, 0 otherwise	.15	.21	.11	.22	.22
AGE1	= 1 for age 30-44, 0 otherwise	.28	.26	.21	.26	.26
AGE2	= 1 for age 45-59, 0 otherwise	.32	.28	.34	.27	.27
AGE3	= 1 for age above 60, 0 otherwise	.25	.25	.34	.25	.25
COLLEGE	= 1 for having college degree, 0 otherwise	.3	.28	.31	.28	.28
MALE	= 1 for male, 0 otherwise	.45	.48	.43	.48	.48
BLACK	= 1 for African American, 0 otherwise	.08	.12	.07	.13	.13

of the same population as the panel, the analyst should impute missing values in $\mathbf{X}$, for example using some MAR model such as a Bayesian bootstrap (Rubin, 1981).

We repeated the simulation study of Section 2.3 but allowing $\mathbf{X}$ to be missing for nonrespondents; see Section 2.7 for details. Repeated sampling properties of the inferences for the coefficients in the regression models for Y_1, Y_2, and R continue to be reasonable. The inferences for some coefficients in the models for W_1 and W_2 exhibit low coverage rates. We do not consider this to be a significant concern, as the models for W_1 and W_2 are not of substantive interest—they are only used for sensitivity analysis.

The APYN data file includes survey weights at each wave. For individuals who responded to wave 1, their weights are the product of design-based weights and post-stratification adjustments for unit nonresponse. In general, analysts should ignore such post-stratification adjustments when assessing sensitivity of design-based analyses to nonignorable unit nonresponse. Instead, after imputing the missing values for the nonrespondents in the initial wave, analysts can use the design weights for all n_p records, assuming they are available for the unit nonrespondents. Unfortunately, in the APYN data we do not have the design weights for nonrespondents to the initial wave (or the refreshment sample), so that we are not able to perform traditional design-based analyses after imputing the missing values. Since our goal is to illustrate the sensitivity analysis rather than perform a specific finite population inference, absent the design-weights for nonrespondents we use Bayesian approaches for inference for the panel. We note that, in cases where design weights are available for nonrespondents, any design-based analysis of wave 2 data also should ignore post-stratification adjustments for attrition, since the AN model is used to account for nonignorable attrition.

2.5.2 Results of Sensitivity Analysis

We begin by fitting the AN model using main effects for the demographic variables and assuming MAR nonresponse in both wave one and the refreshment sample. This is equivalent to fitting the model in (2.3) through (2.7) with $\gamma_1 = \rho_1 = 0$. We use Cauchy prior distributions with a scale parameter of 10 for the intercept terms and scale parameter of 2.5 for all other terms (Gelman et al., 2008); we obtained similar results using Cauchy(0, 10) prior distributions on all of the coefficients. To fit the model, we run three MCMC chains from different starting points for 200,000 iterations. We discard the first 20,000 iterations as burn-in and keep every tenth draw. Diagnostics suggest the chains converge with adequate effective sample sizes (at least 500 for each parameter). For each of set of $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{\rho})$, the multivariate potential scale reduction factor is very close to one. We used 200,000 iterations to ensure satisfactory effective sample sizes, particularly for the intercepts.

Table 2.2 summarizes the parameter estimates for the models in (2.3) through (2.5). The coefficient of Y_2 in the model for R is significant, which suggests the attrition is nonignorable. Based on the model with MAR nonresponse, holding all else constant, a panel participant with "a lot" of interest at wave two is less likely to respond in wave two than a disinterested participant. This tracks the disparities in the marginal probabilities of $Y_{2i} = 1$: 65% in the complete-cases in the panel and 72.2% in the refreshment sample. The strongest predictor of interest in wave two is interest in wave one. Additionally, older and college-educated participants are more likely to be interested in the campaign.

The negative coefficient of Y_2 in the attrition model contradicts conventional wisdom that politically interested respondents are *less* likely to attrite (Bartels, 1999).

Table 2.2: Coefficient estimates and 95% posterior intervals for AN model assuming MAR nonresponse in wave one and in refreshment sample ($\gamma_1 = \rho_1 = 0$). Results based on all $n = 4633$ sampled individuals in wave 1 and the refreshment sample. The model for R suggests nonignorable attrition.

Dependent Variable	Y_1	Y_2	R
Intercept	-1.58 (-1.85, -1.32)	-.13 (-.53, .26)	.52 (.07, 1.13)
AGE1	.21 (-.09, .51)	.10 (-.21, .40)	.22 (-.03, .48)
AGE2	.75 (.47, 1.05)	.33 (.03, .64)	.32 (.06, .57)
AGE3	1.21 (.92, 1.50)	1.08 (.73, 1.43)	.47 (.19, .76)
COLLEGE	.11 (-.07, .30)	.67 (.44, .91)	.59 (.39, .79)
MALE	-.04 (-.20, .13)	-.03 (-.23, .17)	.11 (-.05, .27)
BLACK	.73 (.45, 1.02)	.13 (-.25, .53)	-.25 (-.54, .03)
Y_1	–	2.12 (1.78, 2.47)	.41 (.16, .66)
Y_2	–	–	-.82 (-1.65, -.13)

This counterintuitive finding potentially could be an artifact of nonignorable nonresponse in the panel or refreshment sample. For example, suppose politically disinterested individuals refused to respond in the refreshment sample at higher rates than politically interested individuals, given covariates. The resulting over-representation of $Y_{2i} = 1$ in the complete cases in the refreshment sample would show up as nonignorable attrition like that seen in Table 2.2.

We therefore analyze the sensitivity of conclusions to various mechanisms for nonignorable nonresponse in wave one and the refreshment sample. (We thank Chase Harrison and Michael Henderson for providing guidance on response rate disparities among politically interested respondents.) For nonresponse in wave one, we consider $\gamma_1 \in \{\log 0.5, 0, \log(2)\}$. These three settings imply that sampled individuals with "a lot" of interest in wave one have, respectively, half, the same, or double the odds of responding in wave one as other individuals. For nonresponse in the refreshment sample, we consider $\rho_1 \in \{0, \log(2), \log(3)\}$. These values reflect an assumption that politically interested individuals have greater odds of responding to a cross-sectional survey than politically disinterested individuals, as it is well established that people with more interest in the survey topic tend to respond at higher levels than those less interested in the topic (Goyder, 1987; Groves et al., 2000, 2004). This direction could explain the 7.2% disparity in the panel and refreshment sample respondents' marginal percentages of Y_2, whereas assuming politically-interested respondents are less likely to participate in wave 2 only would magnify the apparent bias due to attrition. We consider it unlikely that $\rho_1 > \log(3)$, which would imply that politically interested individuals have more than 3 times higher odds of responding.

To illustrate the approach, we perform a sensitivity analysis for all coefficients in the models in (2.3) through (2.7). Of course, analysts need not perform such extensive analyses and instead can focus on results most relevant to their analysis.

Table 2.3, Table 2.4, and Table 2.5 display results under different settings of the sensitivity parameters, using the Gibbs sampler outlined in Section 2.3. In the model for Y_1 (Table 2.3), the results are insensitive to values of ρ_1, which reflects assumptions about nonresponse in the refreshment sample. This is not the case for γ_1, which reflects assumptions about nonresponse in the first wave. The intercept varies from about -1.3 when interested wave one participants are less likely to respond to -1.8 when interested wave one participants are more likely to respond. Other coefficients, notably the age coefficients, change magnitudes as well, although the general conclusions remain the same across all values of ρ_1.

In the model for Y_2 (Table 2.4), the results are insensitive to the values of γ_1 and, for the most part, to the values of ρ_1 as well. Only the intercept varies substantially with the values of ρ_1. We discuss implications of this for estimating marginal probabilities of Y_2 below.

In the model for R (Table 2.5), the estimates for the coefficient of Y_2 vary widely depending on the sensitivity parameter settings. In particular, when we assume that the nonresponse in the refreshment sample is MAR ($\rho_1 = 0$), the attrition is NMAR with 95% credible intervals for the coefficient of Y_2 including only negative values. When we assume that nonresponse in the refreshment sample is NMAR with $\rho_1 = \log(3)$, we again estimate nonignorable attrition but now these interval estimates include all positive values; that is, a politically interested panel member has greater odds of responding at wave two than a disinterested one. When we assume that $\rho_1 = \log(2)$, the 95% credible intervals for the coefficient of Y_2 include zero, so

Table 2.3: Posterior means and 95% posterior intervals for coefficients in regression of Y_1 on $\mathbf{X}$ for different values of sensitivity parameters. Here, γ_1 is an additional odds of response in the first wave, and ρ_1 is an additional odds of response in the refreshment sample.

(γ_1, ρ_1)	Intercept	AGE1	AGE2	AGE3	COLLEGE	MALE	BLACK
$(0, 0)$	-1.58	.21	.75	1.21	.11	-.04	.73
	(-1.85, -1.32)	(-.09, .51)	(.47, 1.05)	(.92, 1.50)	(-.07,.30)	(-.20, .13)	(.45, 1.02)
$(0, \log 2)$	-1.59	.21	.76	1.21	.11	-.04	.73
	(-1.85, -1.32)	(-.09, .52)	(.47, 1.05)	(.92, 1.51)	(-.07, .29)	(-.20, .13)	(.45,1.01)
$(0, \log 3)$	-1.59	.22	.77	1.22	.11	-.04	.72
	(-1.86, -1.33)	(-.08, .52)	(.49,1.06)	(.94, 1.51)	(-.07, .29)	(-.21,.13)	(.44,1.01)
$(\log 2, 0)$	-1.81	.38	.95	1.32	.12	-.07	.57
	(-2.07,-1.55)	(.08,.68)	(.67,1.24)	(1.03,1.61)	(-.06, .31)	(-.24,.10)	(.29,.84)
$(\log 2, \log 2)$	-1.81	.38	.95	1.32	.12	-.07	.56
	(-2.08, -1.56)	(.09, .68)	(.67, 1.24)	(1.04, 1.62)	(-.06, .30)	(-.24, .10)	(.28, .84)
$(\log 2, \log 3)$	-1.82	.39	.96	1.33	.12	-.07	.56
	(-2.09, -1.56)	(.09, .69)	(.68, 1.25)	(1.05, 1.62)	(-.06, .30)	(-.24, .09)	(.28,.83)
$(\log \frac{1}{2}, 0)$	-1.29	.00	.51	1.04	.10	.02	.90
	(-1.55, -1.03)	(-.29, .30)	(.23, .80)	(.76, 1.33)	(-.08, .27)	(-.15, .19)	(.62, 1.19)
$(\log \frac{1}{2}, \log 2)$	-1.30	.01	.52	1.05	.10	.02	.89
	(-1.56, -1.04)	(-.29,.31)	(.24,.81)	(.77,1.34)	(-.08,.28)	(-.15,.18)	(.61,1.17)
$(\log \frac{1}{2}, \log 3)$	-1.30	.01	.53	1.06	.10	.01	.89
	(-1.57,-1.04)	(-.29,.32)	(.24,.82)	(.77,1.35)	(-.08,.28)	(-.15,.18)	(.61,1.17)

that we would not have evidence to reject MAR attrition. Clearly, the nature of the unit nonresponse affects whether or not we call the attrition nonignorable.

We also estimate $P(Y_2 = 1)$ in the population for each setting of the sensitivity parameters. To do so, we use a Bayesian version of a post-stratified estimator. We write the population probability as

$$\begin{aligned} P(Y_2 = 1) = &\sum_{k=1}^{32} P(Y_2 = 1 | Y_1 = 1, \mathbf{X} = \mathbf{x}_k) P(Y_1 = 1 | \mathbf{X} = \mathbf{x}_k) P(\mathbf{X} = \mathbf{x}_k) \\ &+ \sum_{k=1}^{32} P(Y_2 = 1 | Y_1 = 0, \mathbf{X} = \mathbf{x}_k) P(Y_1 = 0 | \mathbf{X} = \mathbf{x}_k) P(\mathbf{X} = \mathbf{x}_k) \end{aligned} \tag{2.30}$$

where $\mathbf{x}_k$ ranges over the thirty-two possible demographic combinations. We set each $P(\mathbf{X} = \mathbf{x}_k) = \bar{x}_k^{CPS}$. For each scenario, we compute 50,000 draws of $P(Y_2 = 1)$ based on (2.30) and draws from the posterior distributions of the parameters from each model. These $50,000$ draws represent the posterior distribution of $P(Y_2 = 1)$.

Figure 2.6 summarizes the posterior distributions when $\gamma_1 = 0$ and $\rho_1 \in \{\log .75, 0, \log 1.5, \log 2, \log 3\}$. When the nonresponse in the refreshment sample is MAR ($\rho_1 =$

Table 2.4: Posterior means and 95% posterior intervals for coefficients in regression of Y_2 on $(\mathbf{X}, Y_1)$ for different values of sensitivity parameters. Here, γ_1 is an additional odds of response in the first wave, and ρ_1 is an additional odds of response in the refreshment sample.

(γ_1, ρ_1)	Intercept	AGE1	AGE2	AGE3	COLLEGE	MALE	BLACK	Y_1
$(0, 0)$	-.13	.10	.33	1.08	.67	-.03	.13	2.12
	(-.53, .26)	(-.21, .40)	(.03, .64)	(.73, 1.43)	(.44, .91)	(-.23, .17)	(-.25, .53)	(1.78, 2.47)
$(0, \log 2)$	-.63	.17	.44	1.21	.76	-.02	.09	2.18
	(-.99, -.26)	(-.14, .48)	(.14, .75)	(.87, 1.55)	(.54, .99)	(-.22,.19)	(-.31, .50)	(1.86, 2.53)
$(0, \log 3)$	-.90	.19	.49	1.22	.79	.00	.05	2.13
	(-1.24,-.55)	(-.11, .50)	(.18, .78)	(.89, 1.56)	(.57,1.01)	(-.2,.2)	(-.35, .46)	(1.80,2.47)
$(\log 2, 0)$	-.10	.09	.32	1.07	.66	-.03	.15	2.11
	(-.49, .31)	(-.21, .39)	(.02, .63)	(.72,1.42)	(.43,.89)	(-.23, .16)	(-.24,.54)	(1.77,2.45)
$(\log 2, \log 2)$	-.60	.16	.44	1.21	.76	-.02	.10	2.19
	(-.96, -.23)	(-.15, .47)	(.13, .74)	(.87,1.56)	(.53,.98)	(-.22, .19)	(-.30, .51)	(1.86, 2.52)
$(\log 2, \log 3)$	-.87	.19	.48	1.23	.79	.00	.06	2.15
	(-1.21, -.51)	(-.11, .49)	(.18, .78)	(.90, 1.56)	(.57, 1.01)	(-.2, .2)	(-.34, .46)	(1.81, 2.49)
$(\log \frac{1}{2}, 0)$	-.20	.11	.35	1.10	.68	-.03	.11	2.13
	(-.59, .20)	(-.20, .42)	(.05, .67)	(.75, 1.45)	(.45, .91)	(-.23, .16)	(-.28, .51)	(1.80,2.48)
$(\log \frac{1}{2}, \log 2)$	-.69	.17	.46	1.21	.77	-.01	.07	2.18
	(-1.05,-.31)	(-.14,.48)	(.14,.76)	(.86,1.55)	(.55,1.00)	(-.21,.19)	(-.34,.49)	(1.86,2.52)
$(\log \frac{1}{2}, \log 3)$	-.97	.20	.50	1.22	.80	.00	.03	2.10
	(-1.30,-.62)	(-.10,.51)	(.20,.81)	(.89,1.56)	(.58,1.02)	(-.19,.20)	(-.37,.43)	(1.77,2.44)

0), the posterior mean for $P(Y_2 = 1)$ is 0.69. When we set $\rho_1 = \log 1.5$, the credible interval for $P(Y_2 = 1)$ is $(0.60, 0.69)$, which barely includes 0.69. In contrast, setting $\rho_1 = \log 2$ results in an interval of $(0.57, 0.66)$, which no longer includes 0.69. Thus, if it is plausible that politically interested participants in the refreshment sample have at least 1.5 times higher odds of responding to the refreshment sample in wave two than politically disinterested participants, arguably we should not feel comfortable assuming the refreshment sample unit nonresponse is MAR. Going in the other direction, if the sensitivity parameter is set to $\log .75$, the interval is (0.68, 0.76), which includes 0.69. These results suggest that the estimate of political interest in the population at wave 2 is sensitive to nonresponse in the refreshment sample, but the estimate would be significantly different only at rather substantial levels of bias (i.e., when political interested participants have 1.5 higher odds of responding).

We also estimate $P(Y_1 = 1)$ to estimate differences in interest in the presidential candidates from wave one to wave two. We write the population probability as

$$P(Y_1 = 1) = \sum_{k=1}^{32} P(Y_1 = 1 | \mathbf{X} = \mathbf{x}_k) P(\mathbf{X} = \mathbf{x}_k). \tag{2.31}$$

Table 2.5: Posterior means and 95% posterior intervals for coefficients in regression of R on $(\mathbf{X}, Y_1, Y_2)$ for different values of sensitivity parameters. Here, γ_1 is an additional odds of response in the first wave, and ρ_1 is an additional odds of response in the refreshment sample.

(γ_1, ρ_1)	Intercept	AGE1	AGE2	AGE3	COLLEGE	MALE	BLACK	Y_1	Y_2
$(0, 0)$	.52	.22	.32	.47	.59	.11	-.25	.41	-.82
	(.07, 1.13)	(-.03, .48)	(.06, .57)	(.19,.76)	(.39, .79)	(-.05, .27)	(-.54, .03)	(.16, .66)	(-1.65,-.13)
$(0, \log 2)$	.02	.21	.24	.29	.46	.12	-.26	.10	.20
	(-.29, .36)	(-.04, .46)	(-.01, .50)	(-.01, .58)	(.26,.65)	(-.04, .28)	(-.55,.03)	(-.22,.39)	(-.45,.85)
$(0, \log 3)$	-.17	.20	.19	.16	.36	.12	-.27	-.18	.85
	(-.45, .11)	(-.06, .45)	(-.07, .45)	(-.15, .46)	(.15,.57)	(-.04,.29)	(-.56,.02)	(-.57,.17)	(.17,1.62)
$(\log 2, 0)$	.57	.22	.32	.48	.60	.11	-.25	.43	-.90
	(.11,1.23)	(-.03,.48)	(.07,.57)	(.20, .76)	(.4,.8)	(-.05,.27)	(-.54, .03)	(.18,.67)	(-1.75,-.21)
$(\log 2, \log 2)$	.04	.21	.25	.31	.47	.12	-.26	.12	.12
	(-.27, .38)	(-.04, .46)	(.0, .5)	(.02, .59)	(.26, .66)	(-.04, .28)	(-.54, .03)	(-.19,.41)	(-.51, .75)
$(\log 2, \log 3)$	-.15	.20	.20	.18	.38	.12	-.27	-.13	.74
	(-.42, .14)	(-.06, .45)	(-.05, .46)	(-.13, .48)	(.17, .58)	(-.04, .29)	(-.56, .02)	(-.50, .21)	(.08, 1.47)
$(\log \frac{1}{2}, 0)$	.44	.22	.31	.45	.58	.11	-.25	.38	-.69
	(.03, 1.00)	(-.03,.47)	(.06,.56)	(.17, .73)	(.38, .78)	(-.05, .27)	(-.54, .03)	(.13, .63)	(-1.48, -.04)
$(\log \frac{1}{2}, \log 2)$	-.02	.20	.23	.26	.44	.12	-.26	.05	.32
	(-.32,.31)	(-.05,.45)	(-.02,.48)	(-.03,.55)	(.23,.64)	(-.04,.28)	(-.54,.03)	(-.29,.35)	(-.34,1.00)
$(\log \frac{1}{2}, \log 3)$	-.20	.19	.18	.12	.34	.13	-.28	-.25	1.01
	(-.47,.07)	(-.06,.45)	(-.08,.44)	(-.18,.42)	(.13,.54)	(-.03,.29)	(-.57,.02)	(-.64,.10)	(.34,1.76)

When $\gamma_1 = 0$ and $\rho_1 \in \{\log .75, 0, \log 1.5, \log 2, \log 3\}$, the posterior mean is .30 with a 95% credible interval of (.28, .32). This interval does not overlap with the credible intervals for $P(Y_2 = 1)$. If $\rho_1 = 0$ and $\gamma_1 = \log .5$, the estimate of $P(Y_1 = 1)$ is .33 with an interval of $(.31, .35)$. Keeping $\rho_1 = 0$ and setting $\gamma_1 = \log 2$, the estimate of $P(Y_1 = 1)$ is .27 with an interval of $(.26, .29)$. Thus, regardless of the assumptions about unit nonresponse, there clearly was a large uptick in political interest from November 2007 to October 2008.

2.5.3 Using a Prior Distribution on Sensitivity Parameters

As an alternative to fixing the sensitivity parameters at various plausible values, analysts may want a single set of inferences that averages over their prior beliefs about the values of the sensitivity parameters (Molenberghs et al., 2001, 1999). In this section, we illustrate this process for the APYN analysis by constructing and using prior distributions for γ_1 and ρ_1. To do so, we first specify prior distributions on the proportions of nonrespondents who gave “a lot” of thought to the candidates, and convert these beliefs into distributions for γ_1 and ρ_1.

For γ_1, we construct the prior distribution to reflect our belief that nonrespon-

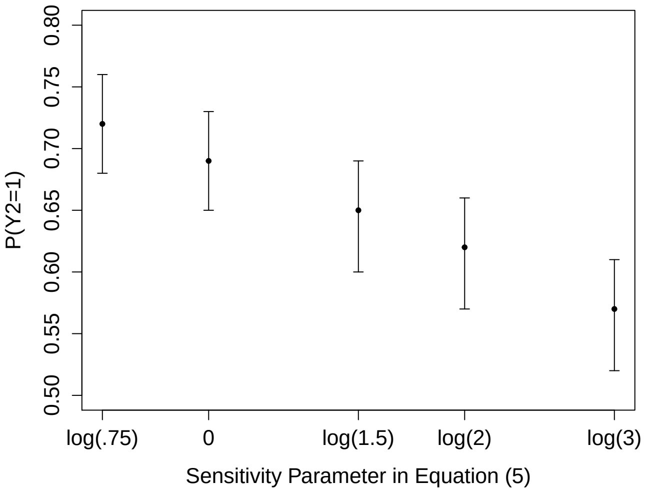


FIGURE 2.6: Posterior means and 95% credible intervals for $P(Y_2 = 1)$ for $\rho_1 \in \{\log .75, 0, \log 1.5, \log 2, \log 3\}$.

dents are not as politically interested as respondents. In wave 1 of the panel, 29.8% of the respondents indicated that they gave "a lot" of thought about the candidates. Thus, we make the 97.5 percentile of the prior distribution for $Pr(Y_1 = 1|W_1 = 0)$ equal to .298. We set the 2.5 percentile of this prior distribution equal to .149, based on consultations with three public opinion experts. Finally, we set the prior distribution for $Pr(Y_1 = 1|W_1 = 0)$ equal to a normal distribution with the matching 95% central interval. One could match to other distributions as well, such as a beta distribution; the normal distribution sufficed to represent our prior beliefs.

We next convert the prior distribution for $Pr(Y_1 = 1|W_1 = 0)$ into a prior distribution for γ_1. To do so, we use the facts that

$$Pr(W_1 = 1 \mid Y_1 = 0) = Pr(Y_1 = 0, W_1 = 1)/Pr(Y_1 = 0) \tag{2.32}$$

$$Pr(W_1 = 1 \mid Y_1 = 1) = Pr(Y_1 = 1, W_1 = 1)/Pr(Y_1 = 1). \tag{2.33}$$

In these equations, we can estimate the two joint probabilities using the empirical percentages from the n_p cases. For any estimate of the marginal probability of Y_1, we can determine the corresponding value of (γ_0, γ_1) by unwinding the selection model for $(W_1 \mid Y_1)$. Ignoring the effects of $\mathbf{X}$ in (2.6) for simplification, given an estimate of $Pr(Y_1 = 1)$ we have

$$logit(Pr(Y_1 = 0, W_1 = 1)/Pr(Y_1 = 0)) = \gamma_0 \tag{2.34}$$

$$logit(Pr(Y_1 = 1, W_1 = 1)/Pr(Y_1 = 1)) = \gamma_0 + \gamma_1. \tag{2.35}$$

Of course, because of the unit nonresponse, we do not know the marginal probability for Y_1. We instead find the sample space for (γ_0, γ_1) by solving (2.34) and (2.35) for all possible empirical percentages of Y_1. For $i = 1, \ldots, 818$ (the total number of nonrespondents in wave 1 of the panel), we compute $p_1^{(i)} = 815 + i$, where 815 is the number of respondents who indicated that they gave "a lot of thought" about the candidates. Then, for each i, we solve

$$logit\left(\frac{Pr(Y_1 = 0, W_1 = 1)}{(n_p - p_1^{(i)})/n_p}\right) = \gamma_0^{(i)} \tag{2.36}$$

$$logit\left(\frac{Pr(Y_1 = 1, W_1 = 1)}{p_1^{(i)}/n_p}\right) = \gamma_0^{(i)} + \gamma_1^{(i)}. \tag{2.37}$$

The collection of $(\gamma_0^{(i)}, \gamma_1^{(i)})$ represents the sample space. We plot the implied CDF of the set of $\gamma_1^{(i)}$ using the probabilities from the normal prior distribution for $Pr(Y_1 = 1|W_1 = 0)$—since γ_1 is effectively a transformed version of $Pr(Y_1 = 1|W_1 = 0)$—and visually match it to a normal CDF function. For γ_1, the resulting normal distribution has mean .39 and standard deviation .22.

For ρ_1, we follow a similar process. For the refreshment sample nonrespondents, we use a normal distribution with central 95% interval from .361 to .722. Here, .722 is the proportion of refreshment sample respondents who had given "a lot" of thought to the candidates. For $i = 1, \ldots, 624$ (the total number of nonrespondents in the refreshment sample), we set $p_2^{(i)} = 333 + i$, and solve the following two equations:

$$logit\left(\frac{Pr(Y_2 = 0, W_2 = 1)}{(n_r - p_2^{(i)})/n_r}\right) = \rho_0^{(i)} \tag{2.38}$$

$$logit\left(\frac{Pr(Y_2 = 1, W_2 = 1)}{p_2^{(i)}/n_r}\right) = \rho_0^{(i)} + \rho_1^{(i)} \tag{2.39}$$

Table 2.6: Coefficient estimates and 95% posterior intervals for AN model with informative prior distributions on sensitivity parameters.

Dependent Variable	Y_1	Y_2	R
Intercept	-1.72 (-2.02, -1.42)	-.53 (-1.08, .07)	.13 (-.33, .79)
AGE1	.31 (.00, .63)	.15 (-.17, .46)	.21 (-.03, .46)
AGE2	.87 (.57,1.18)	.42 (.10, .74)	.26 (.00, .52)
AGE3	1.28 (.98, 1.58)	1.18 (.82, 1.53)	.33 (-.02, .67)
COLLEGE	.12 (-.06,.30)	.74 (.50, .97)	.48 (.23, .72)
MALE	-.06 (-.23, .11)	-.02 (-.22, .18)	.11 (-.04, .28)
BLACK	.63 (.33, .93)	.10 (-.30, .52)	-.25 (-.54, .03)
Y_1	–	2.16 (1.82, 2.51)	.16 (-.35, .56)
Y_2	–	–	.00 (-1.20, 1.18)

As a result, we use a normal distribution with mean .78 and standard deviation .38.

We modify the MCMC algorithm to incorporate these prior distributions. We then run three chains each for 200,000 iterations, discarding the first 20,000 iterations as burn-in and saving every 10th draw. For each of set of ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\tau}$, and $\boldsymbol{\rho}$), the multivariate potential scale reduction factor is less than 1.01. Table 2.6 summarizes the posterior distributions of the regression coefficients. With these prior beliefs on the nonresponse, we find no strong evidence for NMAR attrition, even though the model that assumes MAR nonresponse in wave one and the refreshment sample (see Table 2.2) does suggest NMAR attrition. We also computed posterior inferences for $P(Y_2 = 1)$, obtaining a posterior mean of .63 (95% interval: .54 to .71). Comparing this to Figure 2.6, the posterior mean is in the same range as other settings of sensitivity parameters, and the 95% interval is wider than under fixed settings of the sensitivity parameters.

2.5.4 Model Diagnostics

To check the fit of the models, we follow the advice in Deng et al. (2013) and use posterior predictive diagnostics (Meng, 1994; Gelman et al., 2005; He et al., 2010; Burgette and Reiter, 2010). Using every $100th$ draw from the posterior distributions of the parameters in (2.3) through (2.7), we generate $T = 500$ datasets with new values of (Y_1, Y_2, R, W_1, W_2), holding $\mathbf{X}$ constant. Let $\{D^{(1)}, \ldots, D^{(T)}\}$ be the collection of the T replicated datasets. We then compare statistics of interest in $\{D^{(1)}, \ldots, D^{(T)}\}$ to those in the observed data D. Specifically, suppose that S is some statistic of interest, such as a marginal or conditional probability in our context. For $t = 1, \ldots, T$, let $S_{D^{(t)}}$ be the values of S computed from $D^{(t)}$, and let S_D be the value of S computed from D. We compute the two-sided posterior predictive probability,

$$ppp = (2/T) * \min(\sum_{t=1}^{T} I(S_{D^{(t)}} - S_D > 0), \sum_{t=1}^{T} I(S_{D^{(t)}} - S_D < 0)). \tag{2.40}$$

Small values of *ppp*, for example, less than 5%, suggest that the replicated datasets are systematically different from the observed dataset, with respect to that statistic. When the value of *ppp* is not small, the imputation model generates data that look like the completed data. As statistics S, we use the following quantities: (i) the percentage of cases with $W_1 = 1$, (ii) among all cases with $W_1 = 1$, the percentage of cases with $Y_1 = 1$, (iii) the percentage of cases with $W_2 = 1$, (iv) among all cases with $W_2 = 1$, the percentage of cases with $Y_2 = 1$, and (v) the MLEs of the coefficients in the logistic regression of Y_2 on $(Y_1, \mathbf{X})$, computed with the cases with $W_1 = R = 1$. Each of these quantities is a function of replicated or observed data. We calculate *ppp* values for each S for all settings of the sensitivity parameters in Table 2.3 plus $(0, \log 1.5)$ and $(0, \log .75)$.

The *ppp* values do not suggest obvious problems with model fit. Across all eleven settings, the smallest value of *ppp* is 0.43. About 40% of the *ppp* values are at least 0.90. In other words, the proposed AN model generates replicated data that look like the original data, so that the model does not appear to be implausible. We note that one could examine other diagnostics such as the partial posterior predictive p-value (Bayarri and Berger, 2000, 1998), as additional checks on the fit of the model.

2.6 Concluding Remarks

Most applications of the AN model have ignored nonresponse in the panel and refreshment sample entirely, effectively treating it as missing completely at random. As we have demonstrated, this assumption when unreasonable can lead to substantial bias in estimates. The sensitivity analyses presented here can help analysts investigate how much inferences change with different assumptions about nonignorable nonresponse in either or both of these samples. Similar sensitivity analyses can be used to handle nonignorable dropout when there are multiple waves between the original panel and refreshment sample. To select values of the sensitivity parameters, analysts should consider the extent of selection bias that seems possible for their particular application. For example, analysts can use similar values of ρ_1 and γ_1 when they expect similar reasons explain unit nonresponse in both the initial wave and refreshment sample. Analysts might consider quite different values when this is not expected, for example if the two surveys have different incentive structures. After specifying ranges of interest, analysts can investigate various combinations within the ranges to identify regions for which inferences of interest do not (and do) change meaningfully.

Conceptually, this approach to sensitivity analysis could extend to panels with more waves than considered here. The number of parameters to estimate and the number of sensitivity parameters could become cumbersome, particularly with large dimensions. It may be necessary to use computationally expedient models for the underlying survey data, such as latent class models (Dunson and Xing, 2009; Kunihama and Dunson, 2013), as proposed and implemented by Si et al. (2015c) for two wave surveys with many categorical variables. It also may be necessary to use simplifying assumptions about the sensitivity parameters, for example, set sensitivity parameters equal across waves. Another possibility is to characterize the survey variables with some low rank representation, and perform sensitivity analysis using that representation. Developing sensitivity analysis for surveys with many waves and refreshment samples is an area for future research.

2.7 Supplementary simulations

In this section, we present results from the simulation with $\mathbf{X}$ imputed and the simulation where the data generation and imputation model do not match.

2.7.1 Results of Simulation with $\mathbf{X}$ Imputed

As noted in Section 2.5.1, we ran 100 simulations where $\mathbf{X}$ is not known. Specifically, we generate data as in the simulation in Section 2.3. We then remove $\mathbf{X}$ for the nonrespondents in the panel and refreshment sample, and impute them using the algorithm described in Section 2.5.1.

The results are displayed in Figure 2.7. Repeated sampling properties of the inferences for the coefficients in the regression models for Y_1, Y_2, and R continue to be reasonable. However, the inferences for the coefficients in the models for W_1 and W_2 exhibit low coverage rates. This arises primarily from underestimation of variances rather than from bias. The approach for imputation of X does not fully reflect the uncertainty in the distribution of $\mathbf{X}$, both when completing the samples and when matching to estimated totals. We leave the development of more appropriate imputation approaches for $\mathbf{X}$ to future research, since the central focus of this chapter is on sensitivity analysis for time varying Y values rather than imputation of missing $\mathbf{X}$ data.

2.7.2 Results of Simulation with Different Data and Imputation Models

As discussed in Section 2.3, we investigated how the sensitivity analysis would perform when the underlying models are misspecified. We generated data from robit regression models with one degree of freedom using the same parameter values as in Section 2.3. Specifically, for $i = 1, \ldots, 10000$ cases in each simulated panel and

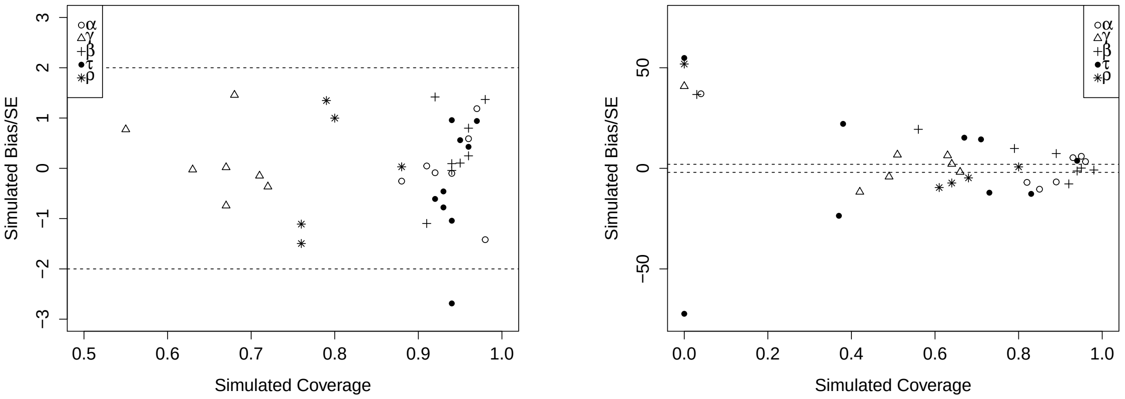


FIGURE 2.7: Simulated coverage versus simulated Bias/SE for different values of sensitivity parameters when values of $\mathbf{X}$ missing due to unit nonresponse are imputed using the algorithm in Section 2.5.1 of the main text. Left panel has sensitivity parameters set to the true values, and right panel has sensitivity parameters set to zero.

$i = 1, \ldots, 5000$ cases in each simulated refreshment sample, we generate data from

$$\mu_{1i} = \alpha_0 + \alpha_X \mathbf{X}_i, \quad Y_{1i} = ((\mu_{1i} + t^*_{1i}) > 0) \tag{2.41}$$

$$\mu_{2i} = \beta_0 + \beta_1 Y_{1i} + \beta_X \mathbf{X}_i, \quad Y_{2i} = ((\mu_{2i} + t^*_{2i}) > 0) \tag{2.42}$$

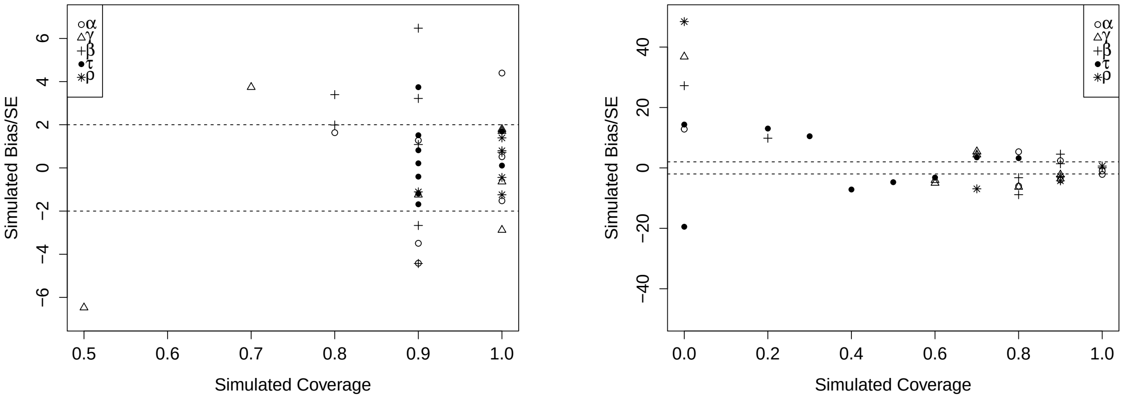
$$\mu_{3i} = \tau_0 + \tau_1 Y_{1i} + \tau_2 Y_{2i} + \tau_X \mathbf{X}_i, \quad R_i = ((\mu_{3i} + t^*_{3i}) > 0) \tag{2.43}$$

$$\mu_{4i} = \gamma_0 + \gamma_1 Y_{1i} + \gamma_X \mathbf{X}_i, \quad W_{1i} = ((\mu_{4i} + t^*_{4i}) > 0) \tag{2.44}$$

$$\mu_{5i} = \rho_0 + \rho_1 Y_{2i} + \rho_X \mathbf{X}_i, \quad W_{2i} = ((\mu_{5i} + t^*_{5i}) > 0). \tag{2.45}$$

Here, each $(t^*_{1i}, \ldots, t^*_{5i})$ are five independent draws from a t-distribution with one degree of freedom. Clearly, this data generation model is not identical to the sequential regressions used in Section 2.3. However, we fit the (misspecified) sequential logistic regression models to perform the sensitivity analysis.

We create ten independent replications of the simulation design. The results are displayed in Figure 2.8. When we set the sensitivity parameters to the true values used in Section 2.3, inferences continue to be more reliable than when we set the sensitivity parameters to zero.



(a) Simulated coverage versus simulated Bias/SE with sensitivity parameters set to true values.

(b) Simulated coverage versus simulated Bias/SE with sensitivity parameters set to 0.

Figure 2.8: Results of 10 replications of the robit model simulations. Data are generated from the robit models, but the sensitivity analysis is done using the sequential logistic regression in (3)–(7).

3

Incorporating Marginal Prior Information in Latent Class Models

The presentation closely follows the work of Schifeling and Reiter (2016).

3.1 Introduction

Mixtures of products of multinomial distributions, also known as latent class models (Goodman, 1974), are used to model multivariate categorical data in many areas of application, including, for example, genomics (Dunson and Xing, 2009), marketing (Kamakura and Wedel, 1997), and political science (Si et al., 2015c). They also serve as engines for multiple imputation of missing data (Vermunt et al., 2008; Gebregziabher and DeSantis, 2010; Si and Reiter, 2013; Manrique-Vallier and Reiter, 2014b). The defining feature of latent class models is an assumption of latent conditional independence: within any class the variables follow independent multinomial distributions. This conditional independence makes latent class models particularly useful for contingency tables with large numbers of cells, as the models can capture complex dependence structures automatically. Bayesian versions of latent class models can be efficiently estimated with MCMC algorithms (Ishwaran and James, 2001; Dunson and Xing, 2009; Jain and Neal, 2004).

In many settings amenable to latent class modeling, the analyst may have informative prior beliefs about the distributions of subsets of the variables. For example, the analyst may know with high precision the distributions of demographic variables from external sources, such as censuses or large national surveys. This information could come in the form of joint distributions (e.g., the probabilities of all combinations of gender and race), conditional distributions (e.g., the probabilities of all combinations of race given gender), or univariate marginal distributions (e.g., the probabilities for all combinations of race and all combinations of gender, separately). It is not obvious how to incorporate such prior information in Bayesian latent class models, because the implied marginal probabilities are tensor products. One approach is the marginally specified prior distribution of Kessler et al. (2014). However, as Kessler et al. (2014) admit, the approximations in this approach can be computationally expensive to implement.

In this chapter, we propose a simple yet highly flexible method for incorporating prior information in Bayesian latent class models. The basic idea is to append synthetic observations to the original data such that (i) the empirical distributions of the desired margins match those in the prior beliefs, and (ii) the values for the remaining variables are left completely missing. For example, to add prior information reflecting that 50% of individuals are female, we can append hypothetical records with only gender recorded and all other variables missing, ensuring that half the augmented records have female for gender. The number of added records is a function of the desired level of prior precision: increasing numbers of records implies increasing certainty in the prior marginal probabilities. After adding the hypothetical records, we estimate the latent class model on the concatenated data with MCMC algorithms. For margins with values in the augmented records, the posterior distribution of the corresponding marginal probabilities is pulled toward the empirical distributions in the augmented records. However, adding the augmented data does not distort conditional distributions of the remaining variables (given the variables with augmented data), since by design the augmented data do not offer information about these conditional distributions. Indeed, as we illustrate, because of this feature the augmented records can be leveraged to correct estimates of the joint distribution of all variables for informative sampling.

Our approach to expressing informative prior distributions is related to the approaches suggested in Greenland (2007) and Kunihama and Dunson (2013). Greenland (2007) adds synthetic records to encode a prior distribution for relative risks, and Kunihama and Dunson (2013) represent prior information by generating pseudo-records with values for all variables using pre-specified, generalized linear models. Unlike these methods, by adding partially complete records our approach allows analysts to encode prior information for arbitrary sets of margins.

The remainder of the chapter is organized as follows. In Section 3.2, we briefly review the particular latent class model that we use, which is a truncated version of the Dirichlet process mixture of product multinomials model (DPMPM) developed by Dunson and Xing (2009). In Section 3.3, we present results of simulations illustrating the augmented record approach, including a discussion of how many records to add. In Section 3.4, we describe how augmented records can be used to account for disproportionate sampling rates in stratified simple random samples. In Section 3.5, we conclude with a brief discussion of other applications of the augmented record approach.

3.2 Review of the DPMPM

In describing the DPMPM, we closely follow the presentation in Si and Reiter (2013). Suppose the data comprise n individuals measured on p categorical variables. Let X_{ij} be the value of variable j for individual i, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let

$X_i = (X_{i1}, \ldots, X_{ip})$. Without loss of generality, we assume that the possible values of X_{ij} are in $\{1, \ldots, d_j\}$, where $d_j \geqslant 2$ is the total number of categories for variable j. Let D be the contingency table formed from all levels of all p variables, so that D has $d = d_1 \times d_2 \times \cdots \times d_p$ cells. We denote each cell in D as $(c_1, \ldots, c_p)$, where each $c_j \in \{1, \ldots, d_j\}$. For all cells in D, let $\theta_{c_1,\ldots,c_p} = \Pr(X_{i1} = c_1, \ldots, X_{ip} = c_p)$ be the probability that individual i is in cell $(c_1, \ldots, c_p)$. We require the $\sum_D \theta_{c_1,\ldots,c_p} = 1$. Let $\theta = \{\theta_{c_1,\ldots,c_p} : c_j \in (1, \ldots, d_j), j = 1 \ldots, p\}$ be the collection of all d cell probabilities.

We suppose that each individual i belongs to exactly one of H^* latent classes. For $i = 1, \ldots, n$, let $z_i \in \{1, \ldots, H^*\}$ indicate the class of individual i, and let $\pi_h = \Pr(z_i = h)$. We assume that $\pi = (\pi_1, \ldots, \pi_{H*})$ is the same for all individuals. Within any class, we suppose that each of the p variables independently follows a class-specific multinomial distribution. This implies that individuals in the same latent class have the same cell probabilities. For any value x, let $\phi_{hjx} = \Pr(X_{ij} = x \mid z_i = h)$ be the probability of $X_{ij} = x$ given that individual i is in class h. Let $\phi = \{\phi_{hjx} : x = 1, \ldots, d_j, j = 1, \ldots, p, h = 1, \ldots, H^*\}$ be the collection of all ϕ_{hjx}. For prior distributions on π and ϕ, we use the truncated stick breaking representation of Sethuraman (1994).

Putting it all together, we have

$$X_{ij}|z_i, \phi \sim \text{Categorical}(\phi_{z_ij1}, \ldots, \phi_{z_ijd_j}) \quad \text{for all } i, j \tag{3.1}$$

$$z_i|\pi \sim \text{Categorical}(\pi_1, \ldots, \pi_{H*}) \quad \text{for all } i \tag{3.2}$$

$$\pi_h = V_h \prod_{g<h}(1 - V_g) \quad \text{for } h = 1, \ldots, H^* \tag{3.3}$$

$$V_h \sim \text{Beta}(1, \alpha) \quad \text{for } h = 1, \ldots, H^* - 1, \quad V_{H*} = 1 \tag{3.4}$$

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha) \tag{3.5}$$

$$\phi_{hj} = (\phi_{hj1}, \ldots, \phi_{hjd_j}) \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}) \tag{3.6}$$

where the Gamma distribution has mean a_α/b_α.

We set $a_{j1} = \cdots = a_{jd_j} = 1$ for all j to correspond to uniform distributions. Following Dunson and Xing (2009) and Si and Reiter (2013), we set $(a_\alpha = .25, b_\alpha = .25)$, which represents a small prior sample size and hence vague specification for the Gamma distribution. In practice, we find these specifications allow the data to dominate the prior distribution. The posterior distribution of all parameters can be estimated using a blocked Gibbs sampler (Ishwaran and James, 2001; Si and Reiter, 2013).

We recommend making H^* as large as possible while still offering fast computation. Using an initial proposal for H^*, say $H^* = 30$, analysts can examine the posterior distributions of the sampled number of unique classes across MCMC iterations to diagnose if H^* is large enough. Significant posterior mass at a number of classes equal to H^* suggests that more classes be added. We note that one can use other MCMC algorithms to estimate the posterior distribution that avoid truncation, for example a slice sampler (Walker, 2007; Dunson and Xing, 2009; Kalli et al., 2009) or an exact blocked sampler (Papaspiliopoulos, 2008).

From 3.1 and 3.2, we can see that the probability of any cell $(c_1, \ldots, c_p) \in D$ can be expressed as

$$\theta_{c_1,\ldots,c_p} = \sum_{h=1}^{H^*} \pi_h \prod_{j=1}^{p} \phi_{hjc_j}. \tag{3.7}$$

Marginal probabilities are computed similarly, taking the product only over the values of j in the margin of interest. This expression reveals the challenge in specifying informative prior distributions for margins in θ: one has to influence both ϕ and π. One possibility is to fix $(a_{j1}, \ldots, a_{jd_j})$ to correspond to the desired prior probabilities with a very large prior sample size that dominates n—this would force the posterior marginal probability to equal the prior marginal probability for variable j. However, this could severely constrain the ability of the model to capture relationships among the other variables, since the prior distribution would encourage the latent classes to be comprised of cases with empirical distributions that match the prior distribution.

3.3 Adding marginal information to the DPMPM model

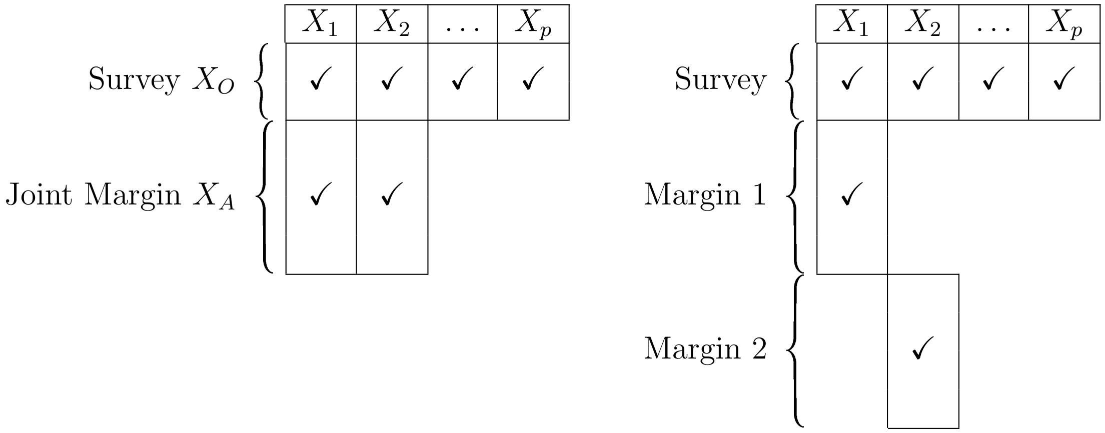
We now turn to the augmented records approach to incorporating prior information about marginal probabilities in the DPMPM model. Let A index the set of variables for which we have informative prior beliefs. Suppose that we create n_A cases to append to the original data. Let X_A include the hypothetically recorded data for the variables in A for the n_A augmented cases; data for all variables not in A are left missing for these cases. Let X_O include all the data for the n cases collected in the sample, and let $X_{obs} = (X_O, X_A)$ be the concatenated data. The exact format of X_{obs} depends on the information in A. When A includes the full joint distribution for $\{X_j : j \in A\}$, the analyst adds X_A as in Figure 3.1a. When A includes only univariate marginal distributions, the analyst adds augmented data comprising only marginal information for each variable $\{X_j : j \in A\}$, as in Figure 3.1b. In the latter case, different augmented sample sizes can be used for each margin depending on the levels of prior precision desired by the analyst.

Let $\Theta = \{z_1, \ldots, z_{n+n_A}, \pi, \alpha, \phi\}$. Treating X_A as if it were data, the likelihood function for the augmented data DPMPM is

$$p(X_{obs}|\Theta) = \left(\prod_{j \in A} \prod_{i=1}^{n+n_A} p(X_{ij}|z_i, \phi)\right) \left(\prod_{j \notin A} \prod_{i=1}^{n} p(X_{ij}|z_i, \phi)\right) \tag{3.8}$$

$$= \left(\prod_{j \in A} \prod_{h=1}^{H^*} \prod_{c_j=1}^{d_j} \phi_{hjc_j}^{\sum_{i=1,z_i=h}^{n+n_A} I(X_{ij}=c_j)}\right) \left(\prod_{j \notin A} \prod_{h=1}^{H^*} \prod_{c_j=1}^{d_j} \phi_{hjc_j}^{\sum_{i=1,z_i=h}^{n} I(X_{ij}=c_j)}\right). \tag{3.9}$$

Using the default prior distributions in (3.3) – (3.6), the posterior distribution of the parameters can be readily estimated with a Gibbs sampler; see Section 3.6 and Appendix A for the full conditionals. The model allows the n_A additional records to



(a) Graphical representation of survey plus a joint distribution margin.

(b) Graphical representation of survey plus two disjoint margins.

FIGURE 3.1: Graphical representations of augmented surveys.

be in any of the latent classes, favoring allocations that best describe X_{obs}.

When n_A is very large, it can be computationally expensive to update each augmented case's z_i one at a time. In many contexts, however, the number of unique combinations in X_A is substantially smaller than n_A; for example, there are two unique combinations when A includes only gender. To update all z_i for the augmented cases, we can compute the conditional probability (given X_A) for each unique combination. We then sample the values of z_i for all augmented records with the same combination at once using a multinomial distribution. When the number of unique combinations in X_A is large, it can be beneficial to update all z_i in parallel. One also can reduce computational burdens by using approximations to the full posterior distribution (e.g., as in Johndrow et al., 2014).

To illustrate the augmented sample approach and the role of n_A, we use three simulation scenarios. In the first scenario, we assume an analyst with very precise (essentially known) estimates of marginal probabilities. Here, we consider prior information comprising a bivariate distribution as in Figure 3.1a and information comprising two univariate margins as in Figure 3.1b. In the second scenario, we assume an analyst with imprecise estimates of marginal probabilities. Here, we only show results for prior information comprising a bivariate distribution. We use a small p in these two scenarios to facilitate repeated sampling studies. In the third scenario, we illustrate the approach for a larger p.

For all simulations, and throughout the remainder of the chapter, we use data from the 2012 American Community Survey (ACS) Public Use Microdata Sample (PUMS) of North Carolina. We include only individuals with age greater than or equal to 18 to avoid structural zeros, i.e., impossible combinations like married five year old. The latent class model from Section 3.2 does not handle structural zeros correctly without adjustments; see Manrique-Vallier and Reiter (2014a) for an approach that does so. The resulting data comprise $N = 76706$ individuals and the variables in Table 3.1. In the following simulations, X_O and the information used to generate X_A both come from this ACS PUMS population. In practice, of course, the survey data in X_O and the marginal information for X_A typically come from different sources.

3.3.1 Scenario 1: Adding known margins

When the analyst knows some marginal probabilities precisely, the analyst should augment the sample with enough records so that $n_A >> n$. As evident from (3.9), doing so ensures that the information about the marginal probabilities in A comes primarily from X_A. The empirical distributions in X_A are constructed to match the known marginal probabilities.

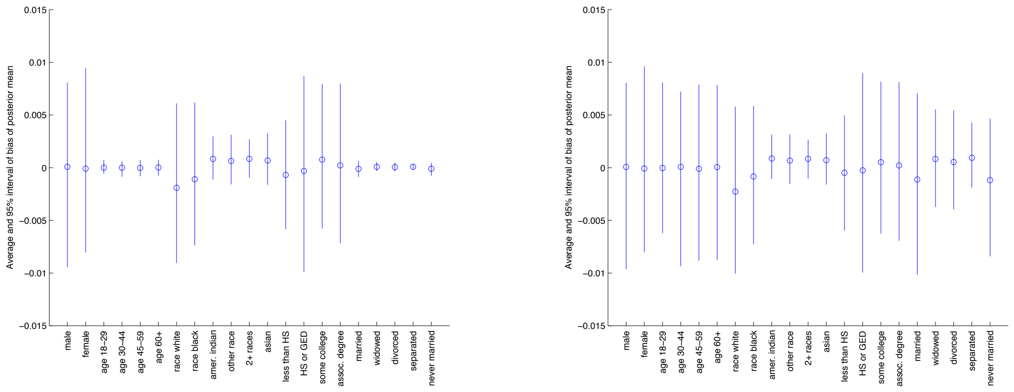
We illustrate this approach using a repeated sampling simulation, treating the N records in the ACS PUMS data as a population. Each X_O comprises $n = 10000$ randomly sampled individuals from the N records in the ACS PUMS. Each record is measured on $p = 5$ variables including gender, age group, race, educational attainment, and marital status, so that the implied contingency table has $d = 960$ cells.

Table 3.1: Subset of variables from ACS PUMS 2012. Categories for age, race, educational attainment, world area of birth, military service, and Hispanic origin have been collapsed from their original number of levels due to insufficient sample sizes.

PUMS variable	Categories
Gender	1=male, 2 = female
Age	1=18-29, 2=30-44, 3=45-59, 4=60+
Recoded detailed race code	1=White alone, 2=Black or African American alone, 3=American Indian alone, 4=other, 5=two or more races, 6=Asian alone
Educational attainment	1=less than high school diploma, 2=high school diploma or GED or alternative credential, 3=some college, 4=associate's degree or higher
Marital status	1=married, 2=widowed, 3=divorced, 4=separated, 5=never married
Language other than English spoken at home	1=yes speaks another language, 2=no speaks only English
World area of birth	1=US state, 2=PR and US island areas, oceania and at sea, 3=Latin America, 4=Asia, 5=Europe, 6=Africa, 7=Northern America
Military service	1=yes active duty at some point, 2=no training for Reserves/National Guard only, 3=no never served in the military
When last worked	1=within the past 12 months, 2=1-5 years ago, 3=over 5 years ago or never worked
Disability recode	1=with a disability, 2=without
Health ins. coverage recode	1=with health insurance coverage, 2=no
Mobility status (lived here 1 year ago)	1=yes same house (non movers), 2=no outside US and PR, 3=no different house in US or PR
School enrollment	1=no has not attended in the last 3 months, 2=yes public school or public college, 3=yes private school or college or home school
Recoded detailed Hispanic origin	1=not Spanish/Hispanic/Latino, 2=Spanish/Hispanic/Latino

We consider an analyst who knows the joint distribution of age group and marital status in the population, which we take from the N records.

We augment each X_O with $n_A = 100000$ synthetic individuals, setting X_A so that the empirical frequencies of the cross tabulations of age group and marital status match those from the known joint marginal probabilities. We run the DPMPM model on X_{obs} with $H^* = 30$, running three MCMC chains each for 50,000 iterations and tossing the first 20,000 as burn-in. We identified this number of MCMC iterates



(a) With $n_A = 100,000$ augmented records.

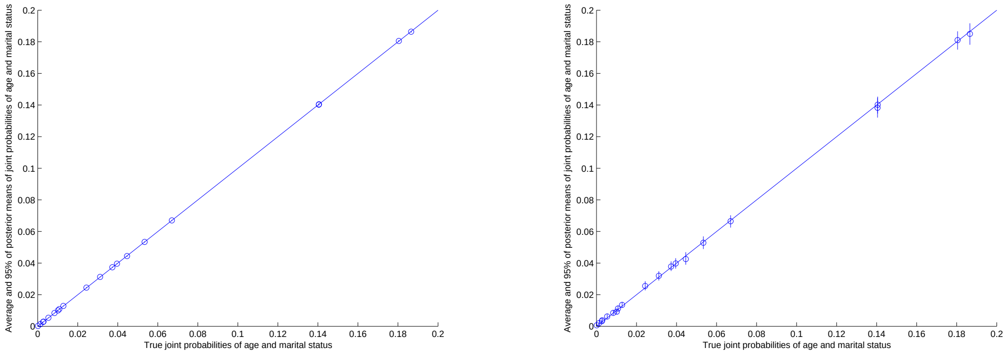
(b) No augmented records.

FIGURE 3.2: Distribution across the 100 simulations of differences in posterior means and corresponding population percentages for all marginal probabilities. Left panel displays results with the augmented joint margin of age group and marital status, and right panel displays results based on collected data only.

as sufficient based on exploratory runs using the diagnostics of Gelman and Rubin (1992) for α and all the univariate marginal probabilities. We repeat the process of generating X_{obs} and fitting the model 100 times. For comparison, we also fit the DPMPM on the 100 sampled X_O without any augmented records.

Figure 3.2 displays how adding X_A affects the estimates of univariate marginal probabilities. After adding the augmented data, the posterior means of the marginal probabilities for age group and marital status are very close to the frequencies in X_A (which equal the population percentages). In contrast, when the DPMPM is estimated using only X_O, the posterior means for the age group and marital status marginal probabilities are substantially more variable. Figure 3.3 shows similar patterns for the joint probabilities of age group and marital status. We note that in Figure 3.2, the posterior means for the marginal probabilities for variables not in A are similar whether or not one adds X_A.

Figure 3.4 displays the posterior means and corresponding population values for all 960 $\theta_{c_1,\ldots,c_p}$. The posterior means are quite similar whether or not one adds X_A. When not using X_A, the average root mean squared error (RMSE) of the posterior



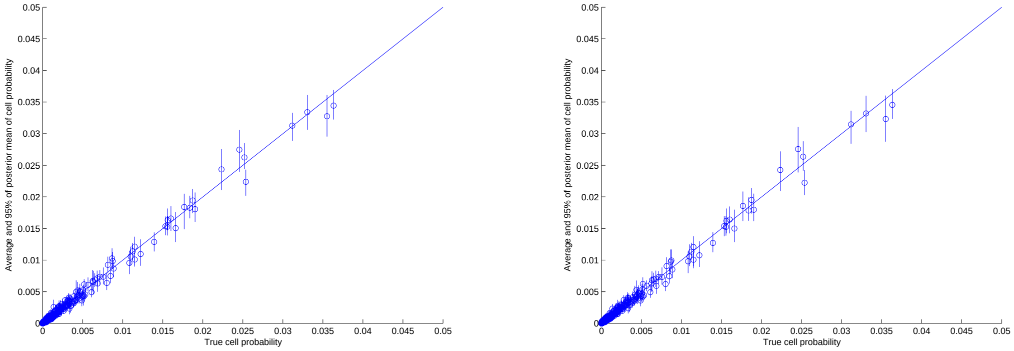
(a) With $n_A = 100,000$ augmented records.

(b) No augmented records.

FIGURE 3.3: Distribution across the 100 simulations of posterior means versus corresponding population percentages for joint distribution of age group and marital status. Left panel displays results with the augmented joint margin of age group and marital status, and right panel displays results based on collected data only.

means is 3.8×10^{-4} with 95% of the RMSEs within $(3.2 \times 10^{-4}, 4.6 \times 10^{-4})$. When using X_A, the average RMSE is 3.8×10^{-4}, with 95% of RMSEs within $(3.1 \times 10^{-4}, 4.5 \times 10^{-4})$. These results indicate that using augmented data to represent prior beliefs on marginal probabilities does not distort other aspects of the posterior distribution of θ.

We also run 100 simulations where the analyst precisely knows the distributions of age group and marital status marginally but not jointly. Here, we add $n_A = 200000$ records as in Figure 3.1b, allocating 100000 to each margin. The results for the 21 univariate marginal probabilities are similar to those in Figure 3.2a, and the results for the 960 cell probabilities are similar to those in Figure 3.4a. When using X_A in this scenario, the average RMSE of the posterior means of the 960 probabilities is 3.9×10^{-4} with 95% of RMSEs within $(3.3 \times 10^{-4}, 4.7 \times 10^{-4})$. These RMSEs are not noticeably different from those in the simulation with known joint age-marital status distribution, although they tend to be slightly higher. However, the posterior probabilities in the joint distribution of age group and marital status exhibit variability that is substantially closer to that seen in Figure 3.3b than in



(a) With $n_A = 100,000$ augmented records.

(b) No augmented records.

FIGURE 3.4: Posterior mean estimates of cell probabilities versus corresponding population values for all 960 cells in the table. Left panel displays results with the augmented joint margin of age group and marital status, and right panel displays results based on collected data only.

Figure 3.3a. This is not surprising, as in this scenario X_A does not add information about the conditional distributions for age group and marital status. For brevity, we do not display the figures here.

3.3.2 Scenario 2: Adding imprecise margins

With imprecise margins, we no longer set $n_A >> n$; instead, we allow n_A essentially to control the prior precision. Suppose that the analyst's prior beliefs for the probabilities in A are centered at some $\theta_A^{(0)}$. When adding augmented data for joint distributions as in Figure 3.1a, analysts can think of n_A as the prior sample size in a Dirichlet distribution with shape parameter $\theta_A^{(0)}$. When adding augmented data for marginal distributions only, analysts specify an augmented sample size for each margin separately. In both cases, the analyst can determine n_A by matching the mean and standard deviation in the prior information (e.g., reported estimates of means and standard errors from national surveys) to the first two moments of Dirichlet distributions. For example, Table 3.2 displays approximate 95% prior intervals on π_j for each possible age group and marital status combination j for various n_A. See

Table 3.2: 95% prior intervals for margins corresponding to different values of n_A.

Age group, marital status	True percent	$n_A = 100000$	$n_A = 10000$	$n_A = 1000$
Age 18-29, married	3.96	(3.84, 4.08)	(3.61, 4.37)	(2.89, 5.24)
Age 18-29, widowed	.01	(.01, .02)	(.00, .06)	(0, .35)
Age 18-29, divorced	.29	(.26, .32)	(.20, .41)	(.11, .85)
Age 18-29, separated	.27	(.24, .31)	(.19, .40)	(.11, .85)
Age 18-29, never married	14.04	(13.82, 14.26)	(13.40, 14.71)	(11.81, 16.04)
Age 30-44, married	14.05	(13.85, 14.26)	(13.37, 14.72)	(11.95, 16.07)
Age 30-44, widowed	.13	(.11, .16)	(.08, .23)	(.03, .54)
Age 30-44, divorced	2.43	(2.34, 2.54)	(2.13, 2.75)	(1.57, 3.55)
Age 30-44, separated	1.01	(.95, 1.08)	(.83, 1.23)	(.51, 1.81)
Age 30-44, never married	5.34	(5.20, 5.46)	(4.90, 5.81)	(4.03, 6.79)
Age 45-59, married	18.04	(17.81, 18.27)	(17.28, 18.72)	(15.48, 20.39)
Age 45-59, widowed	.84	(.79, .89)	(.68, 1.04)	(.39, 1.49)
Age 45-59, divorced	4.47	(4.34, 4.59)	(4.12, 4.89)	(3.28, 5.88)
Age 45-59, separated	1.08	(1.01, 1.14)	(.91, 1.30)	(.63, 1.92)
Age 45-59, never married	3.12	(3.01, 3.23)	(2.79, 3.47)	(2.18, 4.31)
Age 60+, married	18.66	(18.43, 18.90)	(17.84, 19.43)	(16.30, 20.79)
Age 60+, widowed	6.70	(6.56, 6.87)	(6.20, 7.22)	(5.21, 8.21)
Age 60+, divorced	3.73	(3.61, 3.85)	(3.36, 4.09)	(2.68, 5.02)
Age 60+, separated	.53	(.49, .58)	(.42, .69)	(.23, 1.17)
Age 60+, never married	1.28	(1.21, 1.35)	(1.07, 1.51)	(.73, 2.16)

Section 3.7 for further discussion of the reasonableness of interpreting n_A as the prior sample size of a Dirichlet distribution.

To illustrate the incorporation of imprecise marginal information, we modify the simulation from Section 3.3.1. We add prior information on the joint distribution of age group and marital status, using a prior sample size of $n_A = 10000$. Results are summarized in Figure 3.5. As intended, the posterior intervals for the age group and marital status marginal and joint probabilities are wider than those estimated with $n_A = 100000$ yet narrower than those estimated with $n_A = 0$. The average RMSE of the 960 posterior means is again similar to the no-margin and precise-margin cases (the average is 3.7×10^{-4} with 95% of RMSEs between 3.2×10^{-4} and 4.4×10^{-4}).

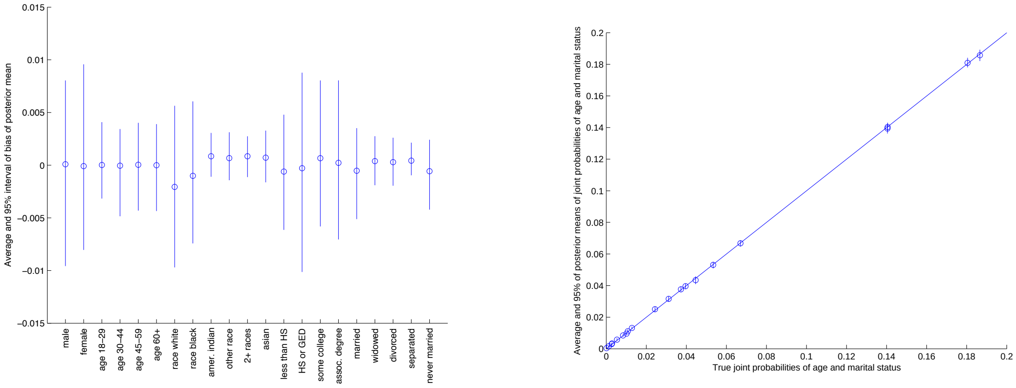


Figure 3.5: Results of 100 simulation runs when $n_A = 10000$. The left panel displays the distribution of differences in posterior means and corresponding population percentages for all univariate distributions, and the right panel displays the posterior means versus the corresponding population percentages for the joint distribution of age group and marital status.

3.3.3 Scenario 3: Adding information with larger p

We now use a random sample of $n = 10000$ records and the $p = 14$ variables in Table 3.1, which correspond to a contingency table with more than 8.7 million cells. We add $n_A = 99991$ (not a multiple of 1000 due to rounding considerations) augmented records with recorded multivariate responses to gender, age group, race, educational attainment, marital status, language other than English, and world area of birth. We construct the augmented data as follows. We compute the population percentage of each combination of these seven variables from the N ACS PUMS cases. For example, the population percentage of people who are male, age 18-29, of white race, have less than a high school diploma, are married, who speak another language other than English, and were born in a US state is .0039%. The cross-tabulation of these seven variables results in 13440 distinct sub-groups, which we allocate to the n_A cases approximately proportional to their population shares.

To investigate the effects of adding prior information, we examine the Cramér's V statistic for every pair of variables j and j'. This measures strength of bivariate associations. Figure 3.6a displays the Cramér's V statistic computed from the N

observations in the ACS PUMS data. Dunson and Xing (2009) define a model-based version of Cramér's V statistic as

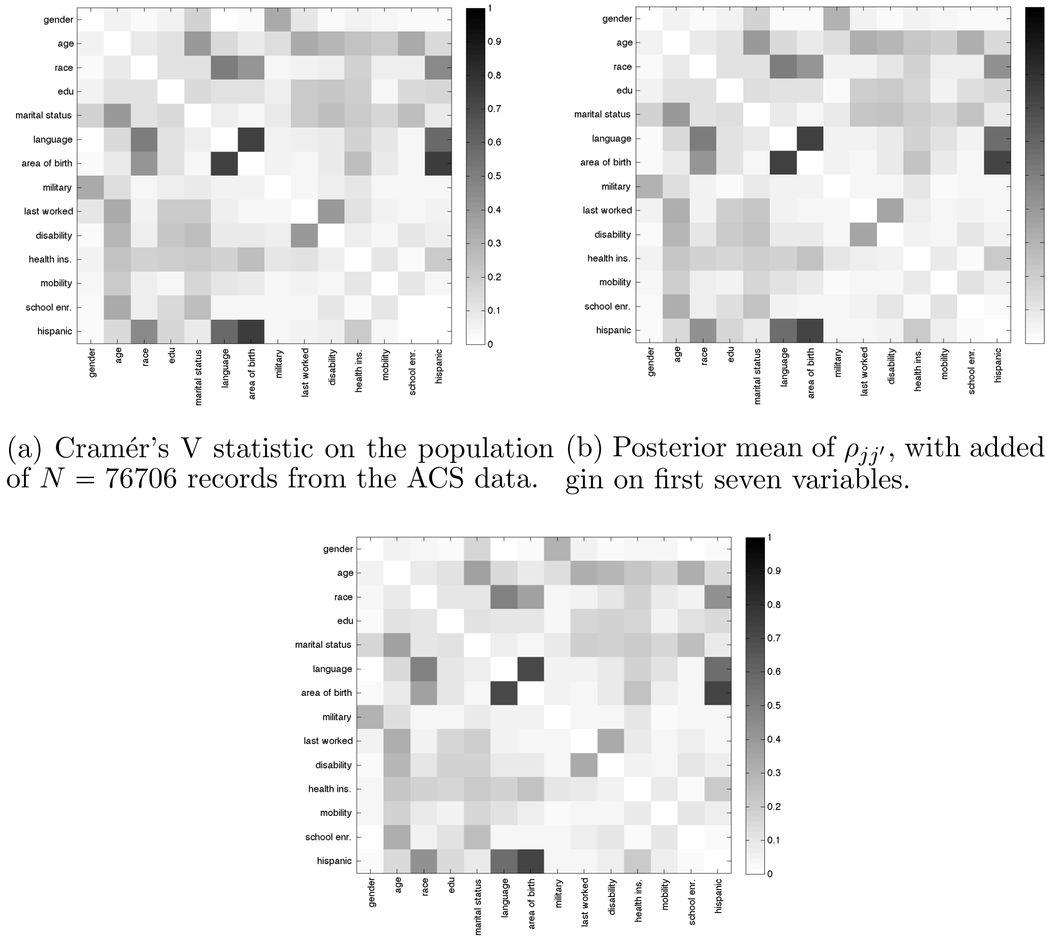
$$\rho^2_{jj'} = \frac{1}{\min\{d_j, d_{j'}\} - 1} \sum_{c_j=1}^{d_j} \sum_{c_{j'}=1}^{d_{j'}} \frac{(\theta_{c_j,c_{j'}} - \theta_{c_j}\theta_{c_{j'}})^2}{\theta_{c_j}\theta_{c_{j'}}} \tag{3.10}$$

where $\theta_{c_j,c_{j'}} = \sum_{h=1}^{H^*} \pi_h \phi_{h,j,c_j} \phi_{h,j',c_{j'}}$ and $\theta_{c_j} = \sum_{h=1}^{H^*} \pi_h \phi_{h,j,c_j}$.

We estimate each $\rho^2_{jj'}$ using the models fit to X_{obs} and only to X_0 using a posterior simulation approach, as done by Dunson and Xing (2009). For each analysis, we run three chains of the MCMC algorithm for 80000 iterations after a burn-in of 20000 iterations, and save every 30^{th} draw. Figures 3.6b and 3.6c display the posterior means of $\rho_{jj'}$ for all pairs of variables. The posterior means of $\rho_{jj'}$ across the variables are similar to the population Cramér's V statistics whether we use X_{obs} or X_O alone. This would not have been the case if, for example, the augmented data encouraged the model to estimate accurately the distribution of the seven variables in the added margin at the expense of the remaining variables. Put another way, the fit based on X_{obs} is not shrunk toward independence relative to the fit based on X_O.

We also consider the joint probabilities in the four-way table involving gender and language spoken other than English (both variables included in A), and school enrollment and Hispanic (not included in A). After adding X_A, the average RMSE of these joint probabilities is .0016 with 95% of values between $(.0009, .0025)$. Without adding X_A, the average RMSE of these probabilities is .0021 with 95% of values between $(.0010, .0036)$. Thus, even though the school enrollment and Hispanic variables are not included in X_A, using the informative prior distribution improves the estimates of the joint probabilities in this four-way table.

Finally, we note that we ran four additional simulations and got similar results for the model-based Cramér's V statistic and the four-way table.



(a) Cramér's V statistic on the population of $N = 76706$ records from the ACS data.

(b) Posterior mean of $\rho_{jj'}$, with added margin on first seven variables.

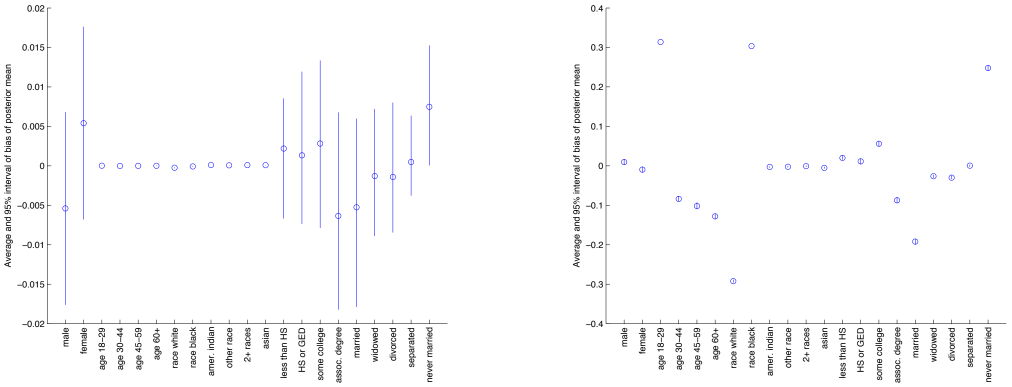
(c) Posterior mean of $\rho_{jj'}$ with no added margin.

FIGURE 3.6: The top figure shows Cramér's V statistic on the population data. The bottom figures show the model-based Cramér's V statistic on the sample of $n = 10000$ records from the ACS data, with and without augmented records.

3.4 Using augmented records to account for stratified sampling

The DPMPM and other Bayesian latent class models effectively treat X_O as coming from a simple random sample. When this is not the case, these and other joint models can result in unreliable inferences about population parameters. In this section, we illustrate how augmented data can be used to adjust for unequal probabilities of selection resulting from stratified random sampling.

We again treat the ACS PUMS data as the population, and use the same $p = 5$ variables as in the simulation in Section 3.3.1. We sample $n = 10000$ records comprising simple random samples of 2500 records from each of four strata, namely African Americans aged 18 to 29, African Americans over age 30, non-African Americans aged 18 to 29, and non-African Americans over age 30. The population shares of the four strata are, in order, 4.2%, 15.3%, 14.4%, and 66.1%. Thus, the stratified sample greatly over-represents younger African Americans and greatly under-represents older non-African Americans. Not surprisingly, when we fit the DPMPM model without correcting for the stratification, the resulting estimates of marginal probabilities are badly biased, as illustrated in Figure 3.7b.



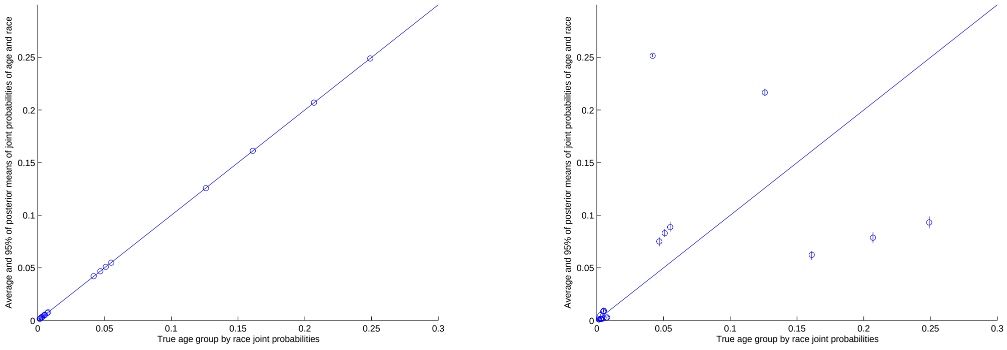
(a) With $n_A = 90000$ augmented records.

(b) No augmented records.

FIGURE 3.7: Difference in posterior means and population quantities for marginal probabilities in the stratified sampling simulation. The left panel fits the model after adding $n_A = 90000$ samples, the right panel fits the DPMPM without any adjustment for stratified sampling. The scales of the vertical axes differ in the two displays to improve interpretation in each display.

In many stratified sampling contexts, the population shares of the strata, and hence of the variables defining the stratification, are known and available for analysis. This suggests that we can treat the known shares as precise prior information and use the techniques of Section 3.3.1. Specifically, we can create augmented records so that the distributions of the stratification variables in the concatenated data match the known population shares. We set n_A large enough that X_{obs} (including X_O) is centered at the population distribution of the stratification variables with negligible variance. Alternatively, when N is not too large and finite population corrections matter, we can set $n_A = N - n$ and choose X_A so that the distribution of X_{obs} exactly matches that in the population.

We run 100 simulations as follows. For each stratified sample of size 10000, we generate $n_A = 90000$ records so that the distribution of age group and race in the concatenated data closely matches the known population shares, leaving all other variables missing in X_A. We assume the analyst knows the joint distribution of all race and all age group combinations, not just the four probabilities used in the stratification. As shown in Figure 3.7a, the DPMPM estimated on X_{obs} results in accurate estimates of the marginal probabilities. This is also the case for the joint distribution of age group and race, as shown in Figure 3.8, and for the 960 cell probabilities, as shown in Figure 3.9.

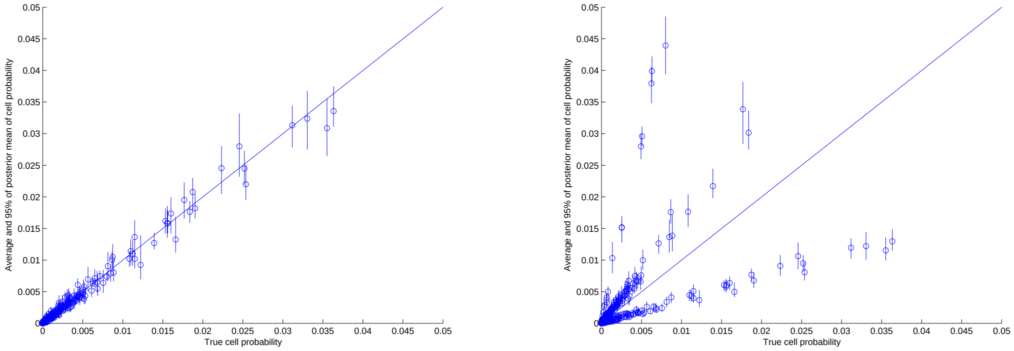


(a) With $n_A = 90000$ augmented records.

(b) No augmented records.

FIGURE 3.8: Posterior mean estimates versus corresponding population values of age group by race joint probabilities. The left panel fits the model with $n_A = 90000$ augmented samples. Across all 100 simulations, all 24 95% credible intervals contain the true joint probability. The right panel fits the model without adjusting for the stratified sampling.

We now offer some intuition on how augmented records can adjust estimated joint



(a) With $n_A = 90000$ augmented records.

(b) No augmented records.

FIGURE 3.9: Posterior mean estimates of cell probabilities versus corresponding population values for all 960 cells in the table. Left panel displays results with $n_A = 90000$ added samples to correct for stratification, and right panel displays results with no added samples.

distributions for stratified sampling. When stratifying on A, by design X_O is not sampled from the population marginal distribution of A. However, because units are collected within strata using simple random samples (this is the standard stratified sampling design), X_O is sampled from the population conditional distribution of $\{X_j : j \notin A\}$ given X_A. Since for large n the DPMPM can accurately estimate the distribution of the generative process for X_O, the DPMPM estimated with only X_O inaccurately estimates the marginal distribution of variables in A, but it should accurately estimate the conditional distribution of $\{X_j : j \notin A\}$ given X_A. Since X_A provides information only about the marginal distribution of A, the DPMPM estimated with X_{obs} still should accurately estimate the conditional distributions of $\{X_j : j \notin A\}$ given X_A. However, X_A encourages the DPMPM to estimate the marginal distributions of A accurately. Fusing the accurate estimates of the marginals of A and conditionals given A results in accurate estimates of the joint distribution. We note that the intuition above assumes that X_O includes all variables used in stratification; otherwise, the conditional distributions implied by the DPMPM are likely to be inaccurate.

The augmented records approach can be further understood using the framework put forth by Kunihama et al. (2014). To adjust the DPMPM for stratified sampling, Kunihama et al. (2014) suggest re-weighting the DPMPM mixture components according to their estimated population shares. The estimated shares are derived from sums of the survey weights of the records in X_O. Augmented records serve a similar function: like estimated shares, they increase or decrease the DPMPM mixture weights to reflect the population distribution of A. The DPMPM tends to assign augmented cases to components occupied by observed cases with similar values of A. These augmented records should not change the distributions of A (or the other variables) within components. Rather, they adjust the mixture weights, as the shares of the components reflect the shares of each combination of A in X_{obs}.

Sometimes the available stratum information is coarser than the corresponding variables used in the analysis; for example, the analyst knows the true proportion of African Americans of age 30 and up from metadata about the survey design, but does not know the breakdown of age 30-44 African Americans, age 45-59 African Americans, and age 60 and up African Americans. In this case, the analyst can construct X_A to match the known percentages at the available coarse scale, and allocate the within-stratum records to match additional prior beliefs about the finer-scale variables. The analyst can create different versions of X_A to reflect different assumptions about the within-stratum allocations, and estimate the DPMPM model on each of the augmented margins as a sensitivity analysis.

3.5 Concluding Remarks

The simulation results presented here suggest that using augmented data is a flexible and convenient way to incorporate prior information about marginal probabilities in latent class models. Augmented categorical cases also could be used to represent prior information on marginal probabilities in other types of mixture models, including models for mixed scale data (e.g., Zhou et al., 2014; Dunson and Bhattacharya, 2011; Wade et al., 2011). The same strategy applies—add n_A cases to reflect prior beliefs about marginal probabilities—with appropriate adjustments to the full conditional distributions. The mixture component indicators for the augmented records can be updated in batch, using only X_A to determine the component probabilities.

The general augmented data approach can be adapted to represent prior information about distributions of continuous variables. For example, to represent the prior belief that the marginal distribution of some continuous variable follows a distribution f, analysts can augment the data with $n_A >> n$ cases drawn from f. Alternatively, analysts can make n_A small to represent relatively weak prior beliefs about the distribution. Analysts can calibrate n_A by examining the properties of summary quantities, such as moments and quantiles, over repeated draws of n_A values of X_A from the prior distribution. This approach could be used to adjust inferences for probability proportional to size samples. The analyst augments X_O with X_A generated to reflect the known, or at least accurately estimated, size distribution in A. We note that the computations with continuous data generally are more challenging, since typically the number of unique values of X_A will be close to n_A.

The approach could be applied in contexts with non-exchangeable data as well. For example, when data comprise people nested within households, analysts may have prior information from census counts on the number of individuals per household, and the distributions of gender and race within households. Given a sample X_O of households, analysts could append augmented household records reflecting those prior beliefs, and estimate appropriate joint models that account for the nested structure (Hu, 2015).

The augmented data approach potentially could improve inferences in other contexts as well. For example, many surveys suffer from unit nonresponse that is not missing at random. If the analyst has external information about the marginal distributions of some of the missing variables, she can augment the sample in a manner like the stratified sampling application and estimate the model on the concatenated data. In this way, the analyst can adjust inferences for nonignorable nonresponse (assuming the data for the variables not in the augmented margins are missing at random). A similar approach could help correct inferences (again under certain conditions) made with convenience samples. We plan to investigate these applications in future research.

3.6 Posterior computation with augmented data

We modify the full conditionals detailed in Appendix A to handle the augmented data. For the augmented data, we do not fill in the missing values of the variables not in A, preferring to marginalize over the missing data. It would be straightforward to impute these missing values, as each variable is independent within latent classes.

1. To update z_i for cases in the augmented data, i.e., where $i = n+1, \ldots, n+n_A$, sample from a categorical distribution with

$$p(z_i = h | \{X_{ij} : j \in A\}, \pi, \phi) = \frac{\pi_h \prod_{j \in A} \phi_{hjX_{ij}}}{\sum_{k=1}^{H*} \pi_k \prod_{j \in A} \phi_{kjX_{ij}}}. \qquad (3.11)$$

   This can be done efficiently by sampling values of z for the recorded combinations in X_A. That is, for each recorded combination in X_A, we compute (3.11) and sample the values of z for all augmented records with that combination using a multinomial distribution.

2. To update ϕ_{hj} for variables with augmented margin, i.e., for $j \in A$, where

$h = 1, \ldots, H^*$, sample from

$$p(\phi_{hj}|X_{obs}, z) = \text{Dirichlet}\left(1 + \sum_{\substack{i=1\\ z_i=h}}^{n+n_A} 1(X_{ij} = 1), \ldots, 1 + \sum_{\substack{i=1\\ z_i=h}}^{n+n_A} 1(X_{ij} = d_j)\right). \tag{3.12}$$

3.7 Interpretation of n_A as prior sample size of Dirichlet distribution

In Section 3.3.2, we suggest thinking of n_A as a prior sample size in a Dirichlet distribution. To illustrate that this is a reasonable interpretation, we now present results of simulation studies in which we approximate the prior distribution on $\phi_{male} = \Pr(\text{gender} = \text{male})$ implied by adding records with only gender recorded.

We take a sample of size $n = 100$ individuals from the PUMS data for whom we observe gender, age group, race, educational attainment, and marital status. We add an augmented sample comprising gender only for $n_A \in \{100, 1000, 10000\}$. We run the DPMPM model on this X_{obs} for $T = 5000$ iterations after the burn-in (also 5000 runs), and save the posterior draws of ϕ_{male}. We repeat the process 100 times.

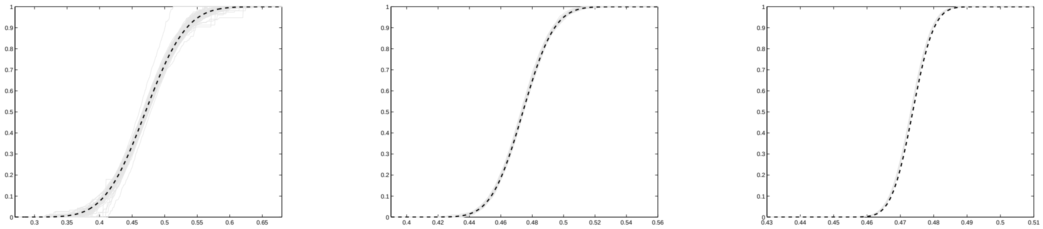
Rearranging Bayes rule, the implied prior distribution of ϕ_{male} given the collected data X_0 is

$$p(\phi_{male}) = \frac{p(\phi_{male}|X_0)p(X_0)}{p(X_0|\phi_{male})}. \tag{3.13}$$

For any simulation run, each of the components on the right hand side of (3.13) can be readily approximated from the converged MCMC samples. Thus, we can approximate $p(\phi_{male})$ along a grid of values from 0 to 1. Let k_1 be the number of males in the sample of $n = 100$ records. At each multiple of .001 between 0 and 1, the approximation is

$$p(\phi_{male} = x) \propto \frac{\frac{1}{T}\sum_{t=1}^{T} I(x - .0005 < \phi_{male}^{(t)} \leqslant x + .0005)}{\frac{n!}{k_1!(n-k_1)!} x^{k_1}(1-x)^{n-k_1}}. \tag{3.14}$$

Figure 3.10 compares the approximated prior cumulative distributions (in gray) to the theoretical Beta(k_2+1, n_A-k_2+1) cumulative distributions (in dashed black line), where k_2 is the number of males in the added margin. For all values of n_A, the Beta distribution is a close match, suggesting that it is reasonable to think of n_A as the prior sample size of a Dirichlet distribution.



(a) $n_A = 100$.

(b) $n_A = 1000$.

(c) $n_A = 10000$.

FIGURE 3.10: Comparison of theoretical Beta CDF to empirical prior CDF under different settings of n_A.

4

Modeling Education Reporting Error by Combining Information from Two Surveys

4.1 Introduction

Survey data are often subject to reporting error. For example, respondents might misunderstand the question or accidentally select the wrong response. The analyst has several options when dealing with such reporting errors. The analyst could ignore the possible reporting errors and treat the values as correct. Or, the analyst could correct the errors using unverifiable assumptions. For example, the analyst could assume that the true value and reported value are conditionally independent given other variables and use this assumption when imputing the true value. The analyst could find a validation sample that includes both the reported, possibly erroneous values along with the true values for a set of individuals. The validation sample allows the analyst to estimate the reporting error mechanism and use this information to impute the true value for the rest of the sample.

Alternately, suppose the analyst has another "gold standard" survey where only the true value is observed. The difference between the gold standard survey and the validation sample is that in the validation sample the true value and reported value are jointly observed for all individuals, whereas in the gold standard survey we assume only the true values are observed. This means in the gold standard survey we do not observe the reporting error mechanism. In this chapter, we develop an approach that allows the analyst to incorporate information from the gold standard survey in modeling reporting error. This research was motivated by a desire to impute true educational attainments of American Community Survey respondents by using information from the National Survey of College Graduates.

More precisely, suppose we have one survey, D_E, in which we observe Z, the reported variable of interest that is subject to reporting error. The "gold standard" survey, D_G, measures the same variable of interest without error, which we denote Y. We assume both surveys collect common demographic variables, which we denote $\boldsymbol{X}$. We would like to use the information in D_G to correct for reporting errors in D_E. However, the values of Y and Z are never jointly observed for any individual. See Figure 4.1 for an illustration of this scenario. This scenario is known as data fusion, when we wish to combine the information from two surveys but some variables are not jointly observed (D'Orazio et al., 2006; Rubin, 1986; Rassler, 2002; Reiter, 2012; Moriarity and Scheuren, 2001).

A common assumption in data fusion is that the variables being fused (Y and Z) are conditionally independent given $\boldsymbol{X}$ (D'Orazio et al., 2006). In some contexts this may be a plausible assumption. In our context, however, when the variables being fused are measuring the same underlying quantity, the conditional independence assumption (CIA) does not make sense. Under the CIA, imputing Y for individuals in D_E would ignore Z, which is essentially Y with some reporting error. That is, the CIA would imply that $p(Y|Z, \boldsymbol{X}) = p(Y|\boldsymbol{X})$. Our goal is to develop a reporting error model for this scenario that does not make the conditional independence assumption.

A related area of research is integrative data analysis (Curran and Hussong,

	$\boldsymbol{X}$	Y	Z
D_E	✓	?	✓
D_G	✓	✓	?

FIGURE 4.1: Graphical representation of data fusion scenario. The D_E variable Z and the D_G variable Y are never jointly observed for any individual.

2009). In particular, Curran and Hussong (2009) describe accounting for between-study measurement heterogeneity in the Cross Study project. In this project, they pool information from three studies on adolescent development. In the framework of a statistical model, they believe there is some underlying latent variable that captures an individual's propensity to experience depressive symptoms, and this latent quantity is measured by the particular questions in each study. This is related to our research question because we also believe Y and Z are measuring the same underlying quantity that is measured differently in both surveys.

The remainder of this chapter is organized as follows. In Section 4.2, we discuss the motivating data example and review work related to data fusion and reporting error modeling. In Section 4.3, we introduce reporting error models that do not assume conditional independence. In Section 4.4, we illustrate various ways of specifying these reporting models and conduct a simulation study to illustrate results. In Section 4.5, we show how to apply the methodology to surveys with complex designs. In Section 4.6, we apply the method to use the 2010 NSCG to impute error-corrected education responses in the 2010 American Community Survey.

4.2 Motivating example and related work

The 1993 National Survey of College Graduates (NSCG) was a resample of individuals who, in the 1990 Census long form, indicated that they had at least a college degree (Fesco et al., 2012). In this special resurvey, it is possible to compare what individuals in the NSCG sample reported as their education level both in the NSCG survey and Census long form. The linked data (NSCG survey linked with the Census long form responses) are available for download from the Inter-University Consortium for Political and Social Research (http://doi.org/10.3886/ICPSR06880.v1) (National Science Foundation, 1993). In later years the NSCG sample also was constructed using previously-reported education level in the sampling design. However, the linked data are not available for public use. Thus, in later years, as is the case in many applications that involve combining information from two surveys, there is no way to link individuals or observations across surveys with public use data. This motivated our investigation into how to utilize the information from a gold standard survey, such as the NSCG, to account for reporting error in more general surveys, such as the American Community Survey. We discuss this application using 2010 data in Section 4.6.

For now we turn back to the 1993 linked NSCG data in which we do observe the reporting error mechanism. Figure 4.2 displays a graphical representation of what we mean by linked data. In D_G we observe both Y and Z for all individuals, so we can investigate the reporting error mechanism $Pr(Z|Y, \boldsymbol{X})$.

Black et al. (2003) analyze the linked data and document the extent of reporting error. In particular, they find that the level of education often is reported higher in the Census than in the NSCG. The linked 1993 NSCG data available from the ICPSR includes individuals that were part of the NSCG sample but reported not having a college degree when asked in the NSCG survey. Of the 214,643 individuals in the

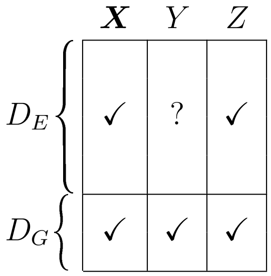
Table 4.1: Unweighted cross-tabulation of NSCG-reported and Census-reported education from the 1993 NSCG linked dataset. We cross-tabulate the data as follows. The Census-reported education is the variable "yearsch". The NSCG-reported education is the maximum degree of the three most recent or highest degrees reported, coded as "ed6c1", "ed6c2", and "ed6c3". We ignore degrees categorized as "other" type and degrees earned in the years 1990-1993. There was a 3-year gap between the Census and the NSCG, and we do not want to consider degrees reported in the NSCG that were earned within this gap.

		Census-reported education				
		BA	MA	Pr.	Ph.D.	*Total*
NSCG-reported education	BA	89580	4109	1241	249	95179
	MA	1218	33928	655	526	36327
	Pr.	382	359	8648	563	9952
	Ph.D.	99	193	452	6726	7470
	Total	91279	38589	10996	8064	148928
	NA: no BA	10150	1792	2040	337	14319
	NA: other	33368	10912	4710	2406	51396

NSCG sample, all of whom had reported at least a college degree in the Census, the NSCG survey found out that over 14,000 individuals did not have a college degree at all (Black et al., 2003). The NSCG-reported education level is the gold standard measurement of education because, as Black et al. (2003) note, the NSCG survey asks many detailed follow up questions about education responses, such as the university where the degree was obtained, the field of degree, and so on.

In spite of the fact that there is nontrivial reporting error, the overwhelming majority of individuals reported consistent levels of education in the Census long form and in the NSCG. We calculate that of the individuals in the NSCG who had at least a college degree at the time of the Census, about 93.3% of them reported an education level consistent with what they reported in the Census. Table 4.1 displays the cross-tabulation of NSCG-reported education and Census-reported education.

The special case of the linked 1993 NSCG data reveals that the majority of NSCG



	$\boldsymbol{X}$	Y	Z
D_E	✓	?	✓
D_G	✓	✓	✓

FIGURE 4.2: The 1993 NSCG sample was selected from all individuals who in the Census reported a college degree. For the individuals in the NSCG sample (D_G), we observe common variables $\boldsymbol{X}$, the NSCG-reported education level Y, and the Census-reported education level Z. In D_E, which comprises individuals *not* in the NSCG sample, we only observe $\boldsymbol{X}$ and Census-reported education Z.

individuals reported consistent levels of education in both the NSCG and Census, so the Census-reported education level should be very useful in predicting true education levels. As noted before, the usual data fusion assumption of conditional independence of Y and Z given $\boldsymbol{X}$ would disregard the information provided by Z when imputing Y.

As an alternative to the conditional independence assumption, D'Orazio et al. (2006) discuss incorporating "auxiliary information" into the data fusion analysis to avoid making the CIA. This auxiliary information can take the form of an additional dataset where the variables being fused are jointly observed, or it can take the form of plausible values of certain inestimable parameters or constraints on parameter values (D'Orazio et al., 2006). In our case, we do not necessarily assume we have auxiliary information about inestimable parameter values; rather, we have information about how the data could be generated. The model we introduce in Section 4.3 can be thought of as constraining certain parameters to be zero in the joint model for $\boldsymbol{X}, Y$ and Z. As another alternative, Fosdick et al. (2015) obtain auxiliary information by collecting new data on the joint distribution of the variables being fused. This approach could be useful when the analyst has additional data related to the reporting error, such as the joint distribution of Y and Z. Related research includes the method presented in Guo and Little (2011), which considers a main study with measurement error in a covariate and a calibration sample where the true covariate and covariate with error are jointly observed.

Kamakura and Wedel (1997) introduce a model-based procedure for data fusion of discrete variables that is based on a finite mixture model. Given two studies, they jointly model the common set of variables as well as the study-specific variables being fused in a finite mixture model. Within each mixture component the variables are locally independent, and so imputing the unobserved responses only depends on the mixture component assignments. Although their model does not explicitly assume conditional independence, there is no additional information about the joint distribution of the study-specific variables. We find in Section 4.4 when using a similar model that the results are similar to the results from the CIA model.

In addition to data fusion, another related area of research is measurement error. In some reporting error contexts the analyst has a "validation sample," which is a sample of individuals for whom both the true value of interest and reported or surrogate value of interest are observed (Pepe, 1992). For example, Schenker and Raghunathan (2007), Schenker et al. (2010), and Raghunathan (2006) discuss combining information from the National Health Interview Survey (NHIS), which has self-reported health data, and the National Health and Nutrition Examination Survey (NHANES), which has both self-reported health data and clinical measures. They construct a reporting error model that predicts clinical outcome given the self-reported data and covariates (Schenker et al., 2010). This is similar to the scenario illustrated in Figure 4.2, where in one survey Y and Z are jointly observed and in the other survey only Z is observed.

Several other reporting error models assume the analyst has data similar to a validation sample. Yucel and Zaslavsky (2005) consider administrative records of reported chemotherapy treatment status in which the true chemotherapy status is obtained for a small subset of patients from their physicians. The reported treatment status is observed for all individuals in the administrative record, and the "true" treatment status is missing for all patients except for those included in the "validation sample." He and Zaslavksy (2009) extend this work to allow for multivariate treatment status.

Zhang et al. (2013) develop a reporting error model that incorporates their knowledge about how the data are generated. In their research, they assume that infant birth weight given true gestational age follows a normal distribution, but gestational age is often missing or misreported (Zhang et al., 2013). Their approach uses a mixture of normal distributions to detect which data points do not belong to the main mixture component for a given gestational age and therefore have an implausible gestational age (Zhang et al., 2013). Our model is similar in the sense that we also incorporate prior knowledge about how the data are likely to be generated.

He et al. (2014) consider the problem where the true outcome of interest, in their case true hospice-use status, is missing for all patients in their sample. Instead, they have two different sources of reported hospice-use status, which both can be subject to misreporting and missing data. Similar to our model specification in Section 4.3, they decompose the joint distribution of true hospice-use status and the two reported statuses into an outcome model (true hospice-use status given covariates) and a reporting model (reported statuses given true status and covariates). They make further assumptions that the two reported statuses are conditionally independent given the true status and covariates, and that the sources may underreport but not overreport (He et al., 2014). We also make untestable modeling assumptions that are reasonable in context and useful in the absence of complete data.

4.3 Model specification

We first introduce notation useful for describing the reporting error model. Let D_G be the dataset that includes the gold-standard variable of interest Y. Let D_E be the dataset that includes the error-prone variable of interest Z. Let D_G and D_E comprise n_G and n_E individuals, respectively, and let $n = n_G + n_E$ be the total size of the combined surveys.

Let i index individuals in the surveys. For each individual i, let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$ be the demographic variables common to both surveys, such as gender and race. We assume these variables have been harmonized across the two surveys (D'Orazio et al., 2006) and that they do not have any reporting error. We assume all variables $(\boldsymbol{X}, Y, Z)$ are categorical, although similar ideas apply for other data types. In particular, let each demographic variable X_j have d_j levels, so that $D = d_1 \cdot d_2 \cdot \ldots \cdot d_p$ is the total number of combinations of demographic variables. Let d_Z be the number of levels of Z, and let d_Y be the number of levels of Y. In the Census/NSCG application, we are only interested in individuals who reported at least a college degree in the Census, resulting in $d_Z = 4$ levels (bachelor's degree, master's degree, professional degree, or PhD). The number of levels of education in Y is 5, with the additional level being "no college degree."

As in Figure 4.1, Y is observed only in D_G and Z is observed only in D_E. We define the variable E to be an indicator of a reporting error, that is, $E_i = 1$ when $Y_i \neq Z_i$ and $E_i = 0$ otherwise. We do not observe E_i for any individuals in D_E or D_G, because we never jointly observe Y_i and Z_i.

We split the full joint distribution into three parts and specify models for each, which we call the true data model, the error model, and the reporting model. For individual i, the full data likelihood is factored as follows (ignoring parameters for simplicity):

$$Pr(\boldsymbol{X}_i = \boldsymbol{x}, Y_i = k, Z_i = l) = Pr(Y_i = k, \boldsymbol{X}_i = \boldsymbol{x}) \\ \times Pr(E_i = e | Y_i = k, \boldsymbol{X}_i = \boldsymbol{x}) \times Pr(Z_i = l | E_i = e, Y_i = k, \boldsymbol{X}_i = \boldsymbol{x}). \quad (4.1)$$

This factorization of the full joint distribution is similar to the one in Yucel and Zaslavsky (2005). As discussed in Yucel and Zaslavsky (2005), one advantage of this factorization is we separate the true data generation process and the error generation process. We learn the true data distribution $(\boldsymbol{X}, Y)$ from the D_G survey data. We can specify different error and reporting models and see how our conclusions vary under different assumptions about the error generation process.

4.3.1 Constraints from data

Before specifying the three parts of the model, we need to determine what information is provided by the observed data shown in Figure 4.1. As in Schifeling et al. (2015) and Si et al. (2015c), the missing data limits the number of parameters we can estimate in the model. In particular, we cannot specify a fully saturated model. Since we assume all the variables are categorical, we can view the complete data (comprising Y, Z, and the p demographic variables in $\boldsymbol{X}$) as a contingency table with $d_Y \cdot d_Z \cdot D$ cells.

The two surveys provide $(d_Y + d_Z - 1) \cdot D$ constraints on this contingency table. We calculate the constraints as follows.

- The combined surveys provide D constraints of the form $Pr(X_1 = c_1, \ldots, X_p = c_p)$ for $c_j \in \{1 : d_j\}$ for $j \in \{1 : p\}$.

- The NSCG data provide $(d_Y - 1) \cdot D$ constraints of the form $Pr(Y = k | X_1 = c_1, \ldots, X_p = c_p)$ for $k = 1, \ldots, d_Y - 1$ and for $c_j \in \{1 : d_j\}$ for $j \in \{1 : p\}$.

- The Census data provide $(d_Z - 1) \cdot D$ constraints of the form $Pr(Z = l | X_1 = c_1, \ldots, X_p = c_p)$ for $l = 1, \ldots, d_Z - 1$ and for $c_j \in \{1 : d_j\}$ for $j \in \{1 : p\}$.

Keeping in mind the number of constraints provided by the data helps guide the model specification.

4.3.2 True data model

The true data model describes the distribution of Y and the demographic variables $X_1, \ldots, X_p$. We assume there is no reporting error in the variables $X_1, \ldots, X_p$ or Y. Any joint model could be used here, such as a loglinear model. Another option is to specify a conditional model for Y given $\boldsymbol{X}$. In our application to the 2010 data, we use a conditional model for $(Y|\boldsymbol{X})$ that utilizes the survey weights in D_G. We discuss this in more detail in later sections. When using a joint model for $(\boldsymbol{X}, Y)$, we suggest using a Dirichlet process mixture of multinomials model (DPMPM) to automatically capture important interactions among the possibly large number of variables (Dunson and Xing, 2009). Intuitively, for the application to the education data, this part of the model learns what true bachelor-degree holders look like, what true master-degree holders look like, and so on.

The DPMPM model is described in detail in Chapter 3. Here, we review it for completeness. Each individual $i = 1, \ldots, n$ belongs to exactly one of H^* latent classes. Let $z_i \in \{1, \ldots, H^*\}$ indicate the latent class assignment of individual i, and let $\pi_h = \Pr(z_i = h)$ for $h = 1, \ldots, H^*$. We assume that $\pi = (\pi_1, \ldots, \pi_{H^*})$ is the same for all individuals. Within any class, the p demographic variables and Y independently follow class-specific multinomial distributions. For any value x, let $\phi_{hjx} = \Pr(X_{ij} = x \mid z_i = h)$ be the probability of $X_{ij} = x$ given that individual i is in class h. Similarly for any value k, we let $\phi_{hyk} = \Pr(Y_i = k \mid z_i = h)$ be the probability of $Y_i = k$ given that individual i is in class h. Let $\phi = \{\phi_{hjx} : x = 1, \ldots, d_j, j = 1, \ldots, p, h = 1, \ldots, H^*\} \cup \{\phi_{hyk} : k = 1, \ldots, d_Y, h = 1, \ldots, H^*\}$ be the collection of all ϕ_{hjx} and ϕ_{hyk}.

The DPMPM model is specified as

$$\begin{aligned}
X_{ij}|z_i,\phi &\sim \text{Categorical}(\phi_{z_ij1},\ldots,\phi_{z_ijd_j}) \quad \text{for all } i, \text{for } j \in \{1:p\} \\
Y_i|z_i,\phi &\sim \text{Categorical}(\phi_{z_iy1},\ldots,\phi_{z_iyd_Y}) \quad \text{for all } i \\
z_i|\pi &\sim \text{Categorical}(\pi_1,\ldots,\pi_{H*}) \quad \text{for all } i \\
\pi_h &= V_h \prod_{g<h}(1-V_g) \quad \text{for } h=1,\ldots,H^* \\
V_h &\sim \text{Beta}(1,\alpha) \quad \text{for } h=1,\ldots,H^*-1, \text{and } V_{H*}=1 \\
\alpha &\sim \text{Gamma}(a_\alpha,b_\alpha) \\
\phi_{hj}=(\phi_{hj1},\ldots,\phi_{hjd_j}) &\sim \text{Dirichlet}(a_{j1},\ldots,a_{jd_j}) \quad \text{for } h=1:H^*, j\in\{1:p\} \\
\phi_{hy}=(\phi_{hy1},\ldots,\phi_{hyd_Y}) &\sim \text{Dirichlet}(a_{y1},\ldots,a_{yd_Y}) \quad \text{for } h=1:H^*. \qquad (4.2)
\end{aligned}$$

Following Si and Reiter (2013), we let $a_\alpha = .25$ and $b_\alpha = .25$. Each element of $(a_{j1},\ldots,a_{jd_j})$ and $(a_{y1},\ldots,a_{yd_Y})$ is set equal to one so that each ϕ_{hj} and ϕ_{hy} has a uniform prior. In the education data, we find that $H^* = 30$ is a sufficient limit to the number of latent classes. To determine if H^* is large enough, the analyst can look at how many of the H^* latent classes are in use, i.e., have individuals assigned to them, at each iteration of the MCMC. If the model is consistently using all or most of the H^* classes, the analyst should increase H^*. If the model consistently uses fewer than H^* classes, H^* is sufficiently large.

4.3.3 Error model

The true data model uses $d_Y \cdot D$ of the available constraints to learn the joint distribution of $(\boldsymbol{X}, Y)$. This leaves $(d_Z - 1)\cdot D$ degrees of freedom to use in the error model and the reporting model. The error and reporting model together specify $Pr(Z_i|Y_i,\boldsymbol{X}_i = \boldsymbol{x})$. For ease of interpretability and computation, we choose to specify the error and reporting models separately as in Manrique-Vallier and Reiter

(2015) and Kim et al. (2015). Intuitively, the error model locates the records for which $Y_i \neq Z_i$, and the reporting model captures the patterns of misreported Z_i. Other possible models include specifying $Pr(Z_i|Y_i, \boldsymbol{X}_i = \boldsymbol{x})$ as a multinomial probit or logit model (Rodriguez, 2007) or using relevant ideas from Agresti (2013) or Dobra et al. (2006) to model the contingency tables defined by $p(Z, Y|\boldsymbol{X} = \boldsymbol{x})$.

There is automatically a reporting error for any individual whose value of Y_i is not in $\{1 : d_Z\}$. For example, in our simulations using the 1993 NSCG data, there is automatically an error when the true education Y is equal to $d_Z + 1$, denoting no college degree, because all individuals in the 1993 NSCG sample reported at least a college degree in the Census. The stochastic part of the error model only applies to individuals who have a true education level of at least a bachelor's degree.

The general error model, for individuals i for whom $Y_i \in \{1 : d_Z\}$, can be specified as

$$Pr(E_i = 1|\boldsymbol{X}_i, Y_i = k) = \Phi(M_i^T \beta). \tag{4.3}$$

$\boldsymbol{M}$ is the model matrix built from $(\boldsymbol{X}_{1:p}, Y)$. The analyst can specify an appropriate prior for the probit parameters β. The analyst changes the error model by changing the model matrix $\boldsymbol{M}$. At the simplest specification, $\boldsymbol{M}$ can be a vector of ones, so that there is a common probability of an error for all individuals. That is,

$$M_i^T \beta = 1 \cdot \beta_0. \tag{4.4}$$

This error model makes sense when the analyst believes the reporting errors in Z are missing completely at random and do not depend on the respondents' demographics or true value of Y. For example, if the reporting errors are simply due to respondents accidentally and randomly selecting the wrong response in the survey, or if all respondents are equally likely to misunderstand the question, this model could be appropriate.

Another model allows the probability of reporting error to depend on some of the demographics, such as gender or race. Intuitively, this model is similar to a missing at random model. For example, if the analyst believes men and women have different probabilities of reporting erroneous responses in Z, the following model allows a different intercept for each gender

$$M_i^T \beta = \beta_0 + \beta_{female} I(X_{i,sex} = \text{F}). \tag{4.5}$$

The analyst might believe that in addition to differences in reporting error by gender, respondents of different races had different probabilities of misunderstanding the question and therefore misreporting their response in Z. Such a model could be specified as

$$M_i^T \beta = \beta_0 + \beta_{female} I(X_{i,sex} = \text{F}) + \sum_{c=2}^{d_{race}} \beta_c I(X_{i,race} = c). \tag{4.6}$$

Alternatively, we can allow the probability of misreporting Z to depend on Y. The interpretation encodes a not missing at random reporting process. For example, when the analyst believes the probability of misreporting Z depends both on the respondent's gender and true value of Y, the following model could be appropriate:

$$M_i^T \beta = \beta_0 + \sum_{k=2}^{d_Z} \beta_k^{(M)} I(Y_i = k, X_{i,sex} = \text{M}) + \sum_{k=1}^{d_Z} \beta_k^{(F)} I(Y_i = k, X_{i,sex} = \text{F}). \tag{4.7}$$

The most detailed specification can include at most $(d_Z - 1) \cdot D$ parameters, comprising of main effects and interaction terms between $\boldsymbol{X}_{1:p}$ and Y. This is the maximum number of parameters that are identifiable from the data, and in practice it may be less, for example, if Y and Z are completely independent of $\boldsymbol{X}$. We recommend specifying fewer parameters that have substantive meaning, rather than specifying the maximum number of parameters that are more difficult to interpret.

4.3.4 Reporting model

If there is no reporting error for individual i, i.e., $E_i = 0$, then we know $Z_i = Y_i$. If there is a reporting error, i.e., $E_i = 1$, then $Z_i \neq Y_i$ and we must model the reported value Z_i.

We could assume that any values of Z_i are equally likely, as in the illustrative example of Manrique-Vallier and Reiter (2015). When Z has d_Z levels, we fix the reporting probabilities at a uniform distribution, that is

$$Pr(Z_i = l|\boldsymbol{X}_i = \boldsymbol{x}, Y_i = k, E_i = 1) = \begin{cases} 1/(d_Z - 1) & \text{if } l \neq k, k \in \{1 : d_z\} \\ 1/d_Z & \text{if } k \notin \{1 : d_Z\} \\ 0 & \text{otherwise.} \end{cases} \tag{4.8}$$

Such a reporting model would make sense if we believe the reporting errors were due to the respondent accidentally selecting the wrong response. While some respondents may have made a simple clerical mistake, Black et al. (2003) found that many respondents in the NSCG had reported higher levels of education in the Census, suggesting (4.8) is not appropriate for the application.

Instead of fixing the reporting probabilities to follow a uniform distribution, we can estimate $p_k(l)$, the probability of reporting $Z = l$ given that $Y = k$, so that the reporting model is

$$(Z_i|Y_i = k, E_i = e) = \begin{cases} \delta_k & \text{if } e = 0 \\ \text{Categorical}(p_k(1), \ldots, p_k(d_Z)) & \text{if } e = 1. \end{cases} \tag{4.9}$$

Note that $p_k(k) = 0$ because if $E = 1$ and $Y = k$ then it is not possible to have $Z = k$. We can use independent Dirichlet priors for each vector of reporting probabilities $(p_k(1), \ldots, p_k(k-1), p_k(k+1), \ldots, p_k(d_Z))$. We can use a uniform Dirichlet prior if we do not have prior information about the distribution of reported levels of Z. Alternatively, the analyst can use a different Dirichlet prior to incorporate prior beliefs about reporting error. For example, the analyst might believe that an individual is more likely to report a level of education close to their true education, so that an individual with no college degree is more likely to report a college degree than a more advanced degree. Or, as we show in Section 4.6, the analyst might have another data source that provides information about reporting error. We use the NSCG data from 1993 to construct a prior for the reporting error in the 2010 data.

The analyst might believe men and women report differently; for example, Black et al. (2003) hypothesized that women were more likely to misunderstand the education question in the Census, as they might report a professional degree when they had no college degree and a professional certification of some kind. In this case, we can parameterize the reporting model such that the reporting probabilities vary with gender, e.g.,

$$(Z_i | \boldsymbol{X}_i, Y_i = k, E_i = e) = \begin{cases} \delta_k & \text{if } e = 0 \\ \text{Cat}(p_{M,k}(1), \dots, p_{M,k}(d_Z)) & \text{if } e = 1, X_{i,sex} = \text{M} \\ \text{Cat}(p_{F,k}(1), \dots, p_{F,k}(d_Z)) & \text{if } e = 1, X_{i,sex} = \text{F}. \end{cases} \tag{4.10}$$

4.3.5 Choosing plausible model specifications

There are many ways to specify the error and reporting models. Error and reporting model specifications that use less than $(d_Z - 1) \cdot D$ degrees of freedom allows us to identify all cells of the contingency table defined by cross-tabulating $\boldsymbol{X}, Y$, and Z. Just as the conditional independence assumption provides an identifiable parameterization of the full contingency table, the error and reporting model can define other identifying assumptions.

When deciding on the parameters to include in a plausible error and reporting models, it is important to consider the specific data set being analyzed and what parameter specifications intuitively make sense, as well as the number of constraints available from the data. For example, if we believe that all the misreported education levels were simply due to respondents accidentally selecting the wrong response in the survey, we could use the simplest error and reporting models, as in (4.4) and (4.8). This model implies that everyone has the same probability of making a mistake, and if they did make a mistake, the education level they reported is completely random.

Instead, we might believe that all respondents were equally likely to misinterpret the question and thought it was asking about their current enrollment in a degree program. Then it would make sense to use a simple error model that assumes everyone has the same probability of misreporting education, as in (4.4), but the education they do report does depend on their true education level, as in (4.9). For example, a PhD student might report a PhD when in fact their highest awarded degree was a master's, or a master's student might report a master's degree when their highest degree was a bachelor's.

We might believe that most individuals either reported their true education level or reported a higher education level, so that the error and reporting models both depend on the respondent's true education level. We might further think that men and women have different probabilities of over-reporting their education, so that gender is also a factor in the models. Then we can specify the error model as in (4.7) and the reporting model as in (4.10). In our application to the education data, the Census has $d_Z = 4$ levels of education and the NSCG has $d_Y = 5$ levels of education; this error model uses 8 parameters and the reporting model uses 22 parameters. We would need to make sure that $(d_Z - 1) \cdot D > 30$ so that all parameters are identifiable.

4.4 Illustration of model specification

In this section we illustrate various ways the error and reporting models can be specified for the 1993 NSCG survey data, and examine results from several different models in an empirical illustration. The demographics we use in this illustration are gender (1 = male, 2 = female) and race, regrouped with six levels (1 = American

Indian, 2 = Asian or Pacific Islander, 3 = Hispanic, 4 = Other, 5 = White, 6 = Black).

4.4.1 Models

We will compare the following models on the 1993 NSCG data.

1. The "MAR fixed" model is specified by (4.5) and a fixed uniform reporting model as in (4.8).

2. The "MAR" model is specified by (4.5) and (4.9).

3. The "NMAR" model is specified by (4.7) and (4.9).

4. The "Gender report" model is specified by the simple error model in (4.4) and a reporting model that depends on gender in (4.10).

In addition to these models, we also model the data using the conditional independence assumption and a mixture model similar to Kamakura and Wedel (1997). The CIA model we estimate is:

$$\begin{aligned}(\boldsymbol{X}_i|\theta) &\sim \text{Categorical}\left(\theta_1, \ldots, \theta_x\right)\\ (Y_i|\boldsymbol{X}_i = \boldsymbol{x}, \phi^G) &\sim \text{Categorical}\left(\phi^G_x\right)\\ (Z_i|\boldsymbol{X}_i = \boldsymbol{x}, \phi^E) &\sim \text{Categorical}\left(\phi^E_x\right). \qquad (4.11)\end{aligned}$$

We use independent uniform Dirichlet priors on the parameters θ, ϕ^G_x for all x, and ϕ^E_x for all x.

The mixture model similar to Kamakura and Wedel (1997) is a DPMPM model in which all variables $\boldsymbol{X}_{1:p}$, Y, and Z are independent within each latent class.

4.4.2 Simulation

To illustrate the implications of these different model specifications, we set up a simple simulation. Although many surveys have complex sampling designs, a point we address in Section 4.5, for now we focus on model specification and so create datasets we can treat as simple random samples.

To do so, we first subset the 1993 NSCG to comprise only the 148928 individuals who completed the interview and the 14319 individuals who reported no bachelor's degree. That is, we ignore the 51396 individuals who were out of scope of the survey for other reasons (see Table 4.1). Next, we sample $n = 100000$ individuals with replacement from the NSCG dataset, where the probability of selecting unit i is proportional to the individual's final survey weight w_i. The first $n_G = 50000$ records are designated as the D_G simple random sample and the remaining $n_E = 50000$ records are designated as the D_E simple random sample. We erase Y in D_E and Z in D_G so that our data are in the form of Figure 4.1.

In the D_G data, we initialize the missing Z to be equal to Y. In the D_E data, we initialize the missing Y to be equal to Z. In practice this seems to be a reasonable starting place for missing data. We run each model's MCMC for 100000 iterations, except for the quickly-converging CIA model which we run for only 20000 iterations. We treat the first 50% of iterations as burn-in.

For each model specification, we obtain $M = 100$ completed D_E surveys by saving imputations of Y throughout the MCMC algorithm (Si and Reiter, 2013; Rubin, 1987). For each of the M completed D_E datasets, we calculate the overall proportion of errors and the proportion of errors by gender and NSCG education. The overall proportion of errors includes individuals with $Y_i =$ no college degree, all of whom have $E_i = 1$ by definition. We combine the point estimates and variances using the `mitools` and `survey` packages in `R`. The results are shown in Table 4.2. In Table 4.3, we show the estimated reporting probabilities under the gender-report model.

Without knowing the linkages, the data do not provide information about which identifying assumption best describes the data; all of the models make untestable

Table 4.2: Summary of results from different model specifications. For each model, we use standard MI combining rules to combine results from the $M = 100$ completed datasets. The table shows the mean and 95% confidence interval. Actual denotes point estimates from the 1993 linked data.

	estimated error rate by group				estimated overall error rate
	Y=BA	Y=MA	Y=Prof.	Y=PhD	
MAR fixed					
Male	.00 (.00, .00)	.00 (.00, .00)	.00 (.00, .00)	.00 (.00, .00)	.08 (.07, .08)
Female	.00 (.00, .00)	.00 (.00, .00)	.00 (.00, .00)	.00 (.00, .01)	
MAR					
Male	.04 (.02, .07)	.04 (.02, .07)	.04 (.01, .08)	.04 (.01, .07)	.14 (.11, .16)
Female	.07 (.04, .09)	.07 (.04, .10)	.06 (.03, .10)	.07 (.02, .11)	
NMAR					
Male	.06 (.03, .09)	.06 (.00, .14)	.08 (.00, .22)	.11 (.00, .27)	.17 (.13, .21)
Female	.08 (.04, .13)	.13 (.00, .29)	.23 (.00, .66)	.59 (.06, 1.00)	
Gender-report					
Male	.05 (.03, .08)	.05 (.03, .08)	.05 (.03, .08)	.06 (.03, .09)	.14 (.12, .16)
Female	.05 (.03, .08)	.06 (.03, .08)	.05 (.02, .09)	.06 (.02, .10)	
CIA					
Male	.36 (.35, .37)	.78 (.77, .80)	.90 (.88, .92)	.95 (.93, .97)	.53 (.53, .54)
Female	.30 (.29, .31)	.77 (.76, .79)	.94 (.92, .97)	.98 (.96, 1.00)	
Kamakura					
Male	.35 (.34, .36)	.78 (.76, .79)	.87 (.85, .89)	.92 (.90, .95)	.53 (.52, .53)
Female	.30 (.29, .31)	.77 (.75, .79)	.91 (.87, .96)	.96 (.93, .99)	
Actual					
Male	.05	.08	.11	.12	.14
Female	.04	.05	.16	.10	

assumptions. The analyst must evaluate the model assumptions and determine which results make the most sense in context. For example, the "MAR fixed" model implies a very small error rate for individuals with at least a college degree, which seems implausible. The CIA and Kamakura models estimate the error rate to be about .53, but the estimated error rate for individuals with professional degrees and PhDs seems very high. The remaining three models (MAR, NMAR, and gender-report)

Table 4.3: Estimated mean and 95% confidence interval of reporting probabilities by gender under the "Gender-report" model.

	Z=BA		Z = MA		Z = Prof.		Z = PhD	
	Actual	Estimated	Actual	Estimated	Actual	Estimated	Actual	Estimated
Y=BA								
Male	-	-	.70	.56 (.29, .82)	.25	.38 (.13, .63)	.05	.06 (.00, .16)
Female	-	-	.75	.39 (.01, .77)	.23	.57 (.20, .95)	.02	.04 (.00, .10)
Y=MA								
Male	.47	.41 (.00, .91)	-	-	.35	.40 (.00, .88)	.18	.19 (.00, .50)
Female	.63	.34 (.00, .85)	-	-	.23	.52 (.00, 1.00)	.14	.14 (.00, .36)
Y=Prof.								
Male	.31	.39 (.00, .90)	.29	.37 (.00, .89)	-	-	.40	.24 (.00, .66)
Female	.42	.35 (.00, .85)	.37	.38 (.00, .87)	-	-	.21	.28 (.00, .74)
Y= PhD								
Male	.11	.33 (.00, .85)	.20	.29 (.00, .73)	.69	.38 (.00, .91)	-	-
Female	.25	.31 (.00, .85)	.32	.34 (.00, .90)	.43	.36 (.00, .95)	-	-
Y=none								
Male	.75	.77 (.62, .92)	.13	.17 (.03, .31)	.10	.03 (.00, .08)	.02	.03 (.00, .06)
Female	.67	.77 (.64, .91)	.09	.14 (.01, .28)	.23	.07 (.00, .16)	.01	.01 (.00, .04)

are fairly similar in estimating an overall error rate (counting individuals with Y_i = no college degree) between .14 and .17. The analyst can consider if the error rates estimated by each model seem plausible. Table 4.2 displays the actual error rates by group that we calculate using the linked data. The actual error rates seem most similar to the error rates estimated by the MAR model or the gender-report model, suggesting that these models are most reasonable in this application.

The gender-report model only has one parameter in the error model, so the estimated error rate by group is fairly consistent with an average estimated error rate around .05 for males and females with at least a college degree. The MAR model estimates a separate probability of error for males and females, with an average estimated error rate around .04 and .07, respectively. The NMAR model specifies a different error rate for each gender and education level. The confidence intervals for the estimated error rates by group are much larger than those in the MAR model. In particular, the 95% confidence interval for the estimated error rate of women with PhDs is (.06, 1.00). The confidence interval for women with professional degrees is also large. There are not many women with professional degrees or PhDs, and so there is not a lot of information from the data for this parameter estimate. The estimated error rates for individuals with bachelor's or master's degrees are not as extreme because there are more individuals with those degrees and thus more information from the data.

We show the reporting model results from the gender-report model in Table 4.3. The model estimates a low probability of reporting a PhD when the true education level is a bachelor's degree. When the true education level is no college degree, the model estimates a large probability of reporting a bachelor's degree and small probabilities of reporting a professional or PhD. Many of the other estimated reporting probabilities have huge confidence intervals, implying that the data do not provide much information.

The gender-report model implies that of individuals with a bachelor's degree who misreport their education level, men are likely to report a master's degree whereas women are more likely to report a professional degree. Comparing this with actual estimates of the reporting probabilities from the 1993 linked data, both men and women with bachelor's degrees who misreport their education tend to report a master's degree with the same probability. Of individuals with no degree, the gender-report model estimates that the proportion of women who report a professional degree is about double the proportion of men who report a professional degree. The same pattern holds in the actual data estimates, but the estimated proportions are higher.

In addition to examining the results in Tables 4.2 and 4.3, we also suggest comparing the distribution of imputed Y in D_E to the observed distribution of Y in D_G. Recall in Section 4.3.1, the D_G data provide constraints of the form $Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})$. The diagnostic we propose aims to check whether the model specification is able to satisfy these constraints. In particular, we use the completed datasets $D_E^{(m)}$, for $m = 1, \ldots, M$, to estimate the probability $Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})$ for each combination of $(\boldsymbol{x}, k)$. We calculate the point estimate and associated variance estimate within each of the M datasets and use multiple imputation combining rules (Rubin, 1987) to find the overall 95% confidence intervals. We calculate similar point estimates from the D_G dataset and calculate how many of the D_G estimates fall within the corresponding 95% confidence intervals from D_E.

Specifically, for $m = 1, \ldots, 100$ and for each $\boldsymbol{x}$ and k, we define $v_{ix} = I(\boldsymbol{X}_i = \boldsymbol{x})$ and $u_{ixk}^{(m)} = I(\boldsymbol{X}_i = \boldsymbol{x}, Y_i^{(m)} = k)$, which varies with each dataset m. The estimated mean and variance using the ratio estimator are

$$\hat{\pi}_{xk}^{(m)} = \bar{u}_{xk}^{(m)}/\bar{v}_x \tag{4.12}$$

$$\widehat{Var}(\hat{\pi}_{xk}^{(m)}) = \frac{1}{n_E \bar{v}_x^2} \frac{\sum_{i=1}^{n_E} (u_{ixk}^{(m)} - \hat{\pi}_{xk}^{(m)} v_{ix})^2}{n_E - 1} \tag{4.13}$$

$$\text{where} \quad \bar{u}_{xk}^{(m)} = (1/n_E) \left(\sum_{i=1}^{n_E} u_{ixk}^{(m)} \right) \tag{4.14}$$

$$\text{and} \quad \bar{v}_x = (1/n_E) \left(\sum_{i=1}^{n_E} v_{ix} \right). \tag{4.15}$$

Equation 4.13 simplifies to $\frac{n_E}{n_E - 1} \frac{\hat{\pi}_{xk}^{(m)}(1 - \hat{\pi}_{xk}^{(m)})}{\sum_{i=1}^{n_E} v_{ix}}$.

We use MI combining rules to calculate a 95% confidence interval for each $\hat{\pi}_{xk}$. We calculate the point estimate of $Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})$ in the D_G dataset, which we denote $\hat{\pi}_{xk}^G = \frac{\sum_{i=1}^{n_G} I(\boldsymbol{X}_i = \boldsymbol{x}, Y_i = k)}{\sum_{i=1}^{n_G} I(\boldsymbol{X}_i = \boldsymbol{x})}$. For our diagnostic we calculate how many of the $\hat{\pi}_{xk}^G$ point estimates fall within 95% confidence interval calculated from D_E.

The demographics in our model illustrations are gender (2 levels) and race (6 levels), so there are 2 × 6 × 5 = 60 such point estimates. In the CIA and Kamakura models, all 60 of the $\hat{\pi}^G_{xk}$ fall within the 95% confidence interval. Although they perform well on this diagnostic, from Table 4.2 we concluded that some of the estimated error rates seemed implausible.

In the MAR model, 58 of the 60 $\hat{\pi}^G_{xk}$ fall within the 95% confidence interval. In the NMAR model, 59 out of 60, and in the gender report model, 58 out of 60. To understand why the error and reporting model specifications do not do as well as the CIA model in this diagnostic, consider the following equation:

$$Pr(Z = l|\boldsymbol{X} = \boldsymbol{x}) = \sum_{k=1}^{5} Pr(Z = l|Y = k, \boldsymbol{X} = \boldsymbol{x})Pr(Y = k|\boldsymbol{X} = \boldsymbol{x}). \tag{4.16}$$

The left hand side of Equation 4.16 is determined by D_E. Information about $Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})$ comes from D_G, but is subject to the model specification $Pr(Z = l|Y = k, \boldsymbol{X} = \boldsymbol{x})$. Ideally, if the model is flexible enough, the estimate of $Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})$ should be close to the estimate from D_G. However, as an extreme example, suppose the model specification $Pr(Z = l|Y = k, \boldsymbol{X} = \boldsymbol{x}) = 1$ if $l = k$ and 0 otherwise. Then the model would impute $Y = Z$ for all individuals in D_E dataset, regardless of the information from D_G. Unless $p(Y|\boldsymbol{X}) = p(Z|\boldsymbol{X})$ for some education levels, this model would do poorly under the diagnostic.

If the diagnostic finds that a small proportion of $\hat{\pi}^G_{xk}$ are contained within the 95% confidence intervals from D_E, it suggests that the analyst may want to change the model, possibly by including more parameters in the error or reporting models. For example, consider the "MAR fixed" model where we fixed the reporting probabilities to be uniform as in Equation 4.8. Under this specification, the diagnostic finds that only 38 of the 60 D_G point estimates $\hat{\pi}^G_{xk}$ are within the 95% confidence intervals from D_E. By adding parameters to the reporting model and therefore allowing the reporting model be more flexible, our "MAR" model performs much better in the diagnostic (58 out of 60). Both the results in Table 4.2 and the diagnostic suggest that the “MAR” model is better than the “MAR fixed” model.

4.5 Accounting for complex survey design

One practical issue we must address when combining information from two surveys is the different sampling designs (Curran and Hussong, 2009; Schenker et al., 2010; Schenker and Raghunathan, 2007). When the datasets both contained the demographic variables that contributed to the sampling design, it is important to include these variables in $\boldsymbol{X}$ so that the error and reporting models are conditional on the design variables (Gelman et al., 2004). As discussed in Chapter 1, including sampling weights in Bayesian modeling is not straightforward.

In Section 4.6, we apply our method to the 2010 NSCG and 2010 ACS (the Census long form questionnaire is no longer in use, so the ACS is the D_E dataset in our application). Part of the 2010 NSCG sample was selected from the 2009 ACS individuals who reported at least a bachelor’s degree. That is, the ACS-reported education, which is outcome variable Z in our reporting error model, was used as a stratification variable in the sampling design (Fesco et al., 2012; Finamore, 2013). Additionally, Z is not released with the 2010 NSCG data. We need some way to account for the complex survey design with only the final survey weights without requiring the design variables.

One approach to accounting for sampling weights, discussed in Chapter 1, is to construct a dataset that can be treated as a simple random sample. When the size of the population N is not that large, it may be possible to use some of these methods to generate several copies of synthetic populations, model each pair of synthetic populations, and combine the results. However, in our application to the 2010 NSCG and ACS, the population is all individuals under the age of 76 in the US who reported a bachelor’s degree or higher in the ACS. We estimate $N \approx 57$ million using the 2010 ACS public-use microdata. Another direction, which we did not explore, is inverse sampling; rather than imputing the entire population, inverse sampling aims to undo complex survey designs by taking repeated subsamples of the complex survey that can be modeled as simple random samples (Rao et al., 2003; Hinkins et al., 1994). This is a potential future direction.

Another approach is to incorporate the weights into the model. However, we have two sets of weights (one from each survey), and it is not clear how to incorporate both sets of weights in the model. D'Orazio et al. (2010, 2012, 2006) compare approaches to incorporating survey weights in a data fusion context, including the file concatenation approach of Rubin (1986). Given surveys A and B, this approach requires that we know the sampling weights w_A and w_B for each unit in both surveys, in order to calculate a new weight w_{AB} for the concatenated surveys (Rubin, 1986). Given the complex survey designs and the fact that we only have the final survey weights, this approach would be difficult to implement.

4.5.1 Proposed method

We propose the following method to account for complex survey design. First, rather than specifying a joint model for $(X_1, \ldots, X_p, Y)$, we specify a conditional model $(Y|X_1, \ldots, X_p)$. Conditional models generally are more amenable to methods for incorporating the survey design than joint models are. Let θ parameterize the conditional model, so that $\theta_{xk} = Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})$.

We use the 2010 NSCG data to inform the distribution of $\theta_x = (\theta_{x1}, \ldots, \theta_{x5})$ for all possible demographic combinations $\boldsymbol{x}$. We can integrate over the missing Z in the NSCG dataset, because the NSCG provides no information about the parameters in the error and reporting models. When imputing missing Y in the ACS, all of the information needed from the NSCG is represented by θ. We factor the posterior as follows, so that at each iteration of the MCMC we begin by drawing θ that does not depend on the ACS data:

$$\begin{aligned} &f(\theta, \beta, \{p_k(l)\}, Y^{(ACS)}, E^{(ACS)} | \boldsymbol{X}^{(ACS)}, Z^{(ACS)}, \boldsymbol{X}^{(NSCG)}, Y^{(NSCG)}) \\ &\quad = f(\theta | \boldsymbol{X}^{(NSCG)}, Y^{(NSCG)}) f(\beta, \{p_k(l)\}, Y^{(ACS)}, E^{(ACS)} | \theta, \boldsymbol{X}^{(ACS)}, Z^{(ACS)}). \end{aligned} \tag{4.17}$$

The goal is to use the NSCG data to specify $f(\theta|\boldsymbol{X}^{(NSCG)}, Y^{(NSCG)})$. First we parameterize $\theta_{xk} = T_{xk}/\sum_{j=1}^{5} T_{xj}$, where T_{xk} is the population count of individuals with $\boldsymbol{X} = \boldsymbol{x}$ and $Y = k$. If we are able to sample T_{xk} for each $\boldsymbol{x}$ and k, then we can compute each θ_{xk} and we have our desired draw of θ. The general roadmap is to specify a model for T_{xk} that is (1) easy to sample from and (2) is always greater than zero.

If the 2010 NSCG dataset included those individuals who were selected to be in the NSCG sample but then turned out not to have a college degree (as in the 1993 dataset), we could use the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) to estimate the population totals T_{xk} for all five values of k, as in (4.18). Although the NSCG survey has a stratified design, the design variables are not all publicly available. Therefore, we simply use the final survey weights, w_i where $i = 1, \ldots, n_{NSCG}$, in our calculations. We assume probability proportional to size sampling with replacement for purposes of variance estimation, which is common in design-based inferences (Lohr, 2010).

For each $\boldsymbol{x}$ and k, we compute the estimated total and associated variance,

$$\hat{T}_{xk} = \sum_{i=1}^{n_G} w_i I(\boldsymbol{X}_i = \boldsymbol{x}, Y_i = k) \tag{4.18}$$

$$\widehat{Var}(\hat{T}_{xk}) = \frac{n_G}{n_G - 1} \sum_{i=1}^{n_G} \left(w_i I(\boldsymbol{X}_i = \boldsymbol{x}, Y_i = k) - \frac{\hat{T}_{xk}}{n_G} \right)^2 . \tag{4.19}$$

For each k and l, with $l \neq k$, we also compute the estimated covariances,

$$\widehat{Cov}(\hat{T}_{xk}, \hat{T}_{xl}) = \frac{n_G}{n_G - 1} \sum_{i=1}^{n_G} \left[\left(w_i I(\boldsymbol{X}_i = \boldsymbol{x}, Y_i = k) - \frac{\hat{T}_{xk}}{n_G} \right) \right.$$
$$\left. \times \left(w_i I(\boldsymbol{X}_i = \boldsymbol{x}, Y_i = l) - \frac{\hat{T}_{xl}}{n_G} \right) \right]. \quad (4.20)$$

To get a draw $(\theta^*_{x1}, \ldots, \theta^*_{x5})$, we could simply draw $(T^*_{x1}, \ldots, T^*_{x5}) \sim N(\hat{T}_x, \hat{\Sigma}(\hat{T}_x))$, with $(\hat{T}_x, \hat{\Sigma}(\hat{T}_x))$ estimated in (4.18) - (4.20). Given this draw of T^*_x, we then let $\theta^*_{xk} = T^*_{xk} / \sum_{j=1}^{5} T^*_{xj}$. This method may work fine if, in practice, the draws of T^*_x never have negative components. If one of the T^*_{xk} is negative, then the corresponding θ^*_{xk} is negative, and is not a valid multinomial probability for the model it parameterizes.

Rather than assuming $T_x = (T_{x1}, \ldots, T_{x5}) \sim N(\hat{T}_x, \hat{\Sigma}(\hat{T}_x))$, which has the possibility of drawing negative values, we assume $T_x \sim \text{Log-Normal}(\mu_x, \tau_x)$. By assuming T_x follows a multivariate log-normal distribution, we ensure all draws of T_x are nonnegative. We can compute estimates of μ_x and τ_x such that the mean of T_x is $\hat{T}_x$ and the covariance of T_x is $\hat{\Sigma}(\hat{T}_x)$ (Tarmast, 2001). We have

$$T_x = (T_{x1}, \ldots, T_{x5}) \sim \text{Log-Normal}(\mu_x, \tau_x) \quad (4.21)$$

$$\tau_x[j, j] = \log \left(1 + \hat{\Sigma}_x[j, j] / (\hat{T}^2_{xj}) \right) \quad (4.22)$$

$$\tau_x[j, i] = \log \left(1 + \hat{\Sigma}_x[j, i] / (\hat{T}_{xj} \cdot \hat{T}_{xi}) \right) \quad (4.23)$$

$$\mu_{xj} = \log(\hat{T}_{xj}) - \tau_x[j, j]/2. \quad (4.24)$$

In our experience the estimated τ_x might not be positive semidefinite, so we calculate $\tilde{\tau}_x$ which is a matrix close to τ_x that is positive semidefinite. We make use of the function “nearestSPD” in Matlab (D’Errico, 2013).

We can use this log-normal distribution to draw $T^*_x \sim \text{Log-Normal}(\mu_{x,}, \tilde{\tau}_x)$, and then compute $\theta^*_{xk} = T^*_{xk} / \sum_{j=1}^{5} T^*_{xj}$. This is both easy to sample from and guarantees that θ^*_x is nonnegative. This approach of drawing a Log-Normal vector and transforming it to the simplex is related to the Logistic-Normal distribution; using results from Aitchison and Shen (1980), if $(T^*_{x1}, \ldots, T^*_{x5}) \sim$ Log-Normal$(\mu_{x,}, \tilde{\tau}_x)$ then $(\theta_{x1}, \ldots, \theta_{x4}) \sim$ Logistic-Normal$(A\mu_x, A\tilde{\tau}_x A^T)$ where where $A = [I_4, -1_4]$. We thank Mauricio Sadinle for his input on this step.

So far we have outlined the roadmap to sample θ_x starting with estimates of the population totals and associated variances from survey D_G with complex sampling design. In our application to the 2010 NSCG data, we have to include an additional step to handle the fact that the 2010 NSCG dataset does not include individuals for whom $Y = 5$ (no college degree). We know that some individuals who reported at least a college degree in the ACS do not actually have a college degree because the ACS estimate of the total number of college graduates is greater than the NSCG estimate.

From the 2010 NSCG we can estimate $\hat{T}_{xk}$ for $k \in \{1 : 4\}$ and the associated covariance matrix $\hat{\Sigma}\left((\hat{T}_{x1}, \ldots, \hat{T}_{x4})\right)$. We are not able to directly estimate the total number of individuals in the NSCG sample who did not have a college degree, $\hat{T}_{x5}$, nor the associated covariances.

We can use the ACS (survey D_E in our application) to estimate $T_{x,total}$, the total count of individuals with $\boldsymbol{X}_i = \boldsymbol{x}$ and reported education $Z_i \in \{1 : 4\}$. This estimated total is useful because $T_{x,total} = T_{x1} + T_{x2} + T_{x3} + T_{x4} + T_{x5}$, and T_{x5} is the quantity we are actually interested in estimating. In particular, for each $\boldsymbol{x}$, we use (4.18) and (4.19) to estimate the total population count $\hat{T}_{x,total}$ and associated variance $\hat{\sigma}^2(\hat{T}_{x,total})$. If each of $(T_{x1}, T_{x2}, T_{x3}, T_{x4}, T_{x5})$ is assumed to be independently normally distributed, then $T_{x,total}$ is also normally distributed, with mean equal to the sum of their estimated means and variance equal to the sum of their estimated variances. It may be possible to solve for $\hat{T}_{x5}$ and $\hat{\sigma}^2(\hat{T}_{x5})$ that satisfies these equations. However, in our case the estimated variance $\hat{\sigma}^2(\hat{T}_{x,total})$ was smaller than the sum of the individual estimated variances. This suggests that T_{x5} has nontrivial covariance with $(T_{x1}, \ldots, T_{x4})$. If the estimate of $\hat{T}_{x5} = \hat{T}_{x,total} - \sum_{j=1}^{4} \hat{T}_{xj}$ is negative, then there is a discrepancy in the datasets which this method will not solve.

For each $\boldsymbol{x}$, we want to use the distribution of $(T_{x1}, \ldots, T_{x4})$ and $T_{x,total}$ to approximate the joint distribution of $(T_{x1}, \ldots, T_{x5})$. We proceed as follows.

1. Sample $T^*_{x,total} \sim \text{Normal}(\hat{T}_{x,total}, \hat{\sigma}^2(\hat{T}_{x,total}))$ using the ACS data.

2. Sample $(T^*_{x1}, \ldots, T^*_{x4}) \sim \text{Normal}\Big((\hat{T}_{x1}, \ldots, \hat{T}_{x4}), \hat{\Sigma}\left((\hat{T}_{x1}, \ldots, \hat{T}_{x4})\right)\Big)$ using the NSCG data.

3. Compute $T^*_{x,5} = T^*_{x,total} - \sum_{i=1}^{4} T^*_{xi}$. Let $T^*_x = (T^*_{x1}, T^*_{x2}, T^*_{x3}, T^*_{x4}, T^*_{x5})$ we a draw from the distribution of $T_x = (T_{x1}, T_{x2}, T_{x3}, T_{x4}, T_{x5})$. In some cases $T^*_{x,5}$ may be negative.

4. Repeat steps (1) - (3) 10000 times.

From these 10000 draws, we compute the mean vector and covariance matrix, which we denote $\hat{T}_x$ and $\hat{\Sigma}(\hat{T}_x)$, respectively. All elements of $\hat{T}_x$ should be nonnegative, which should not be a problem because $\hat{T}_x$ should be very similar to $(\hat{T}_{x1}, \ldots, \hat{T}_{x4}, \hat{T}_{x,total} - \sum_{j=1}^{4} \hat{T}_{xj})$, which we assumed was nonnegative. This is the end of our necessary additional step, and we assume T_x is approximately normally distributed with mean $\hat{T}_x$ and covariance matrix $\hat{\Sigma}(\hat{T}_x)$. We follow the roadmap to constructing a distribution of θ_x as previously described.

1. We assume $T_x = (T_{x1}, \ldots, T_{x5})$ is approximately distributed $\text{Normal}(\hat{T}_x, \hat{\Sigma}(\hat{T}_x))$.

2. We estimate μ_x and τ_x such that T_x is approximately distributed Log-Normal(μ_x, τ_x). By using a Log-Normal distribution instead of a Normal distribution we ensure all T_{xk} are nonnegative.

3. Given a draw T^*_x, we let $\theta^*_{xk} = T^*_{xk} / \sum_{j=1}^{5} T^*_{xj}$.

4.5.2 Justification of proposed method

We justify this approach with an illustration. In the next section we will analyze the 2010 NSCG and 2010 ACS surveys, and the demographics in our model will be gender (2 levels), age group (4 levels), and an indicator for black race (2 levels). There are $2 \times 4 \times 2 = 16$ demographic combinations. Using (4.18), (4.19), and (4.20), we calculate the estimates $\left(\hat{T}_{x,total}, \hat{\sigma}^2(\hat{T}_{x,total})\right)$ from the 2010 ACS and $\left((\hat{T}_{x1}, \hat{T}_{x2}, \hat{T}_{x3}, \hat{T}_{x4}), \hat{\Sigma}((\hat{T}_{x1}, \hat{T}_{x2}, \hat{T}_{x3}, \hat{T}_{x4}))\right)$ from the 2010 NSCG. If the estimate of $\hat{T}_{xk}$ from the NSCG is equal to 0 for any $\boldsymbol{x}$ and k, then we replace it with the value 0.5. Next, as described above, we draw 10000 samples T^*_x for each $\boldsymbol{x}$. We let $\hat{T}_x$ equal the sample mean and $\hat{\Sigma}(\hat{T}_x)$ equal the sample covariance. As in step (2), we compute μ_x and τ_x. For each $\boldsymbol{x}$ we draw $T^{(1)}_x, \ldots, T^{(S)}_x$ where S=10000. Let $\theta^{(1)}, \ldots, \theta^{(S)}$ be the corresponding S draws of θ.

To determine if the draws of θ are from the appropriate distribution, we check the following quantities:

1. For each $\boldsymbol{x}$ and $k \in \{1 : 4\}$, $\bar{\theta}_{xk} = (1/S) \sum_{s=1}^{S} \theta^{(s)}_{xk}$ should be approximately equal to $\hat{T}^{(NSCG)}_{xk} / \hat{T}^{(ACS)}_{x,total}$. For $k = 5$, $\bar{\theta}_{x5}$ should be approximately equal to $(\hat{T}^{(ACS)}_{x,total} - \sum_{j=1}^{4} \hat{T}^{(NSCG)}_{xj}) / \hat{T}^{(ACS)}_{x,total}$. In our simulation, the largest absolute difference between a $\bar{\theta}_{xk}$ and corresponding estimate from the data is less than .004.

2. For each $\boldsymbol{x}$ and $k \in \{1 : 4\}$, the sample mean of $\left(\theta^{(s)}_{xk} / (\theta^{(s)}_{x1} + \theta^{(s)}_{x2} + \theta^{(s)}_{x3} + \theta^{(s)}_{x4})\right)$ should be approximately equal to $\hat{\theta}_{xk}$. $\hat{\theta}_{xk}$ is computed from NSCG survey using (4.25)-(4.27). In our simulation the largest absolute difference is less than .004.

3. For each $\boldsymbol{x}$ and $k \in \{1 : 4\}$, the sample standard deviation of the quantity

$\left(\theta^{(s)}_{xk}/(\theta^{(s)}_{x1}+\theta^{(s)}_{x2}+\theta^{(s)}_{x3}+\theta^{(s)}_{x4})\right)$ should be approximately equal to $\sqrt{\hat{\sigma}^2(\hat{\theta}_{xk})}$, as calculated in (4.28). In our simulations, the largest absolute difference is less than .002. For reference, the estimated standard errors of $\hat{\theta}_{xk}$ range from 0 to about 0.05.

The estimates $\hat{\theta}_{xk}$ and $\hat{\sigma}^2(\hat{\theta}_{xk})$ are computed from the NSCG survey using ratio estimation (Lohr, 2010). For each individual i, define $u_{ixk} = w_i I(\boldsymbol{X}_i = \boldsymbol{x}, Y_i = k)$, and define $v_{ix} = w_i I(\boldsymbol{X}_i = \boldsymbol{x})$. We have

$$\bar{u}_{xk} = (1/n_G)\left(\sum_{i=1}^{n_G} u_{ixk}\right) \tag{4.25}$$

$$\bar{v}_x = (1/n_G)\left(\sum_{i=1}^{n_G} v_{ix}\right) \tag{4.26}$$

$$\hat{\theta}_{xk} = \bar{u}_{xk}/\bar{v}_x \tag{4.27}$$

$$\hat{\sigma}^2(\hat{\theta}_{xk}) = \frac{1}{(n_G)(\bar{v}_x)^2}\frac{\sum_{i=1}^{n_G}(u_{ixk}-\hat{\theta}_{xk}v_{ix})^2}{n_G-1}. \tag{4.28}$$

4.6 Application: 2010 data

We are now ready to apply the methodology to model reporting error in the 2010 American Community Survey (ACS) education variable. We use the public-use microdata for the 2010 NSCG and the 2010 ACS. Recall that although the 1993 NSCG dataset includes the Census-reported education level, the 2010 NSCG does not include the previously-reported education level, so we do not observe the true reporting error. The goal is to impute Y, the gold standard education as reported in the NSCG, for the 2010 ACS individuals whose ACS-reported education Z is at least a bachelor's degree. As a rough way to see if there is reporting error in the ACS, we calculate the weighted count of males with at least college degrees from the ACS and the NSCG. The weighted count from the ACS is larger than the count from the NSCG and suggests that about 10% of males who reported at least a college degree in the ACS do not truly have a college degree. A similar calculation suggests that about 10% of females who reported at least a college degree in the ACS do not have a college degree.

We save $M = 50$ imputations of Y under different models. Using these multiple imputations, we can (1) explore the reporting error mechanism described by each model, and (2) see how sensitive our substantive conclusions are to the different model specifications. In particular, we are interested in the average income by education level and gender, and the number of science and engineering degrees awarded to women at each each education level.

The substantive question about income is motivated by the analysis in Black et al. (2008, 2006) using the 1993 NSCG. Black et al. (2008) look at gender wage disparities for college-educated women of different races (black, Hispanic, Asian, and non-Hispanic white). Black et al. (2006) look at wages of black, Hispanic, and Asian men in comparison to non-Hispanic white men. They find that some of the apparent wage gaps in the data are due to reporting error of education in the Census (Black et al., 2006). A report from the National Research Council (2008) notes that one important reason for the NSCG is to collect data on women and minorities in the science and engineering workforce; this motivates our substantive question about science and engineering degrees awarded to women.

The analysis proceeds as follows. First, we subset the ACS microdata to include only individuals who reported a bachelor's degree or higher and are under the age of 76. The resulting sample size of the ACS is 600150. The demographic variables include gender, age group (24 and younger, 25 - 39, 40 - 54, and 55 and older), and an indicator for black race. The ACS variable names were "sex", "agep", and "racblk". The corresponding variables in the NSCG were "gender", "agegr", and "black". In the NSCG we discarded 38 records that had race suppressed.

Table 4.4: Summary of model specifications for 2010 data application.

	Error model $M_i^T\beta =$	Reporting model $(Z_i \mid Y_i = k, E_i = 1) =$
Model 1	$\beta_1 + \sum_{k=2}^4 \beta_k I(Y_i = k)$	Categorical$(p_k(1), \ldots, p_k(4))$
Model 2	$\beta_1 + \sum_{k=2}^4 \beta_k^{(M)} I(Y_i = k, X_{i,sex} = \text{M})$ $+ \sum_{k=1}^4 \beta_k^{(F)} I(Y_i = k, X_{i,sex} = \text{F})$	Categorical$(p_k(1), \ldots, p_k(4))$
Model 3	$\beta_1 + \sum_{k=2}^4 \beta_k^{(no)} I(Y_i = k, X_{i,black} = \text{no})$ $+ \sum_{k=1}^4 \beta_k^{(yes)} I(Y_i = k, X_{i,black} = \text{yes})$	Categorical$(p_k(1), \ldots, p_k(4))$
Model 4	$\beta_1 + \sum_{k=2}^4 \beta_k^{(M)} I(Y_i = k, X_{i,sex} = \text{M})$ $+ \sum_{k=1}^4 \beta_k^{(F)} I(Y_i = k, X_{i,sex} = \text{F})$	Cat$(p_{M,k}(1), \ldots, p_{M,k}(4))$ if $X_{i,sex} = \text{M}$ Cat$(p_{F,k}(1), \ldots, p_{F,k}(4))$ if $X_{i,sex} =$F
Model 5	$\beta_1 + \sum_{k=2}^4 \beta_k^{(M)} I(Y_i = k, X_{i,sex} = \text{M})$ $+ \sum_{k=1}^4 \beta_k^{(F)} I(Y_i = k, X_{i,sex} = \text{F})$ β prior from 1993 data	Cat$(p_{M,k}(1), \ldots, p_{M,k}(4))$ if $X_{i,sex} = \text{M}$ Cat$(p_{F,k}(1), \ldots, p_{F,k}(4))$ if $X_{i,sex} =$F $p_{M,k(\cdot)}$ and $p_{F,k(\cdot)}$ priors from 1993 data

4.6.1 Models

The complete model specifications we consider are as follows. For all models, the true data model is $Pr(Y = k \mid X_{sex} = x_1, X_{age} = x_2, X_{black} = x_3, \theta) = \theta_{(x_1,x_2,x_3)k}$. We estimate the distribution of θ from the 2010 NSCG as described in Section 4.5. We consider several models for reporting errors, summarized in Table 4.4.

The reporting error in the 1993 NSCG depended on the true level of education. Our simplest model, "model 1," specifies error and reporting models that only depend on Y. The error model is

$$M_i^T\beta = \beta_1 + \sum_{k=2}^{4} \beta_k I(Y_i = k). \tag{4.29}$$

The reporting model is given in (4.9).

The next two models keep the reporting model as in (4.9) but expand the error model. The error model in our "model 2" is specified as in (4.7), so that the probability of a reporting error can vary by gender and Y. "Model 3" specifies an error model that can vary with Y and a demographic indicator for black race,

$$M_i^T\beta = \beta_1 + \sum_{k=2}^{4} \beta_k^{(no)} I(Y_i = k, X_{i,black} = \text{no}) + \sum_{k=1}^{4} \beta_k^{(yes)} I(Y_i = k, X_{i,black} = \text{yes}). \quad (4.30)$$

In "model 4," the error and reporting models both depend on Y and gender; the specifications are as in (4.7) and (4.10).

For "model 5" we use the same specification as "model 4" but the priors for parameters in both the error and reporting models are constructed using the 1993 linked data. The goal of this model is to see what our imputations of Y look like when we incorporate prior information about the reporting error mechanism from the 1993 NSCG.

We thank Dr. Seth Sanders for his helpful input regarding how reporting errors in the 1993 NSCG data are likely to translate to the 2010 data. He recommended that before constructing our priors from the 1993 data, we first remove records from the 1993 NSCG for whom Census-reported education was imputed, because these imputations were inaccurate. The allocation flag variable for Census education indicates that 4828 records had education imputed. Of these 4828 records who were imputed to have at least a college degree in the Census, 1213 did in fact have at least a college degree as reported in the NSCG, and 3615 did not have a college degree. (Other records in the NSCG had imputed Census education, but they were marked out of scope for other reasons and are not included in our analysis.)

We use the remainder of the 1993 NSCG data to construct the prior distributions. The error rates in 2010 are likely to be similar to those in 1993, so we center the prior at the error rate estimate from the 1993 data. More specifically, our prior for each $\beta_k^{(x)}$ as defined in (4.7) is as follows, where the superscript (93) signifies the

1993 NSCG data.

$$\Phi(\beta_k^{(x)}) \sim \text{Beta}(s \times \frac{a}{a+b}, s \times \frac{b}{a+b}) \tag{4.31}$$

$$a = \sum_{i=1}^{n_G^{(93)}} w_i^{(93)} I(Y_i^{(93)} = k, X_{i,sex}^{(93)} = x, E_i^{(93)} = 1) \tag{4.32}$$

$$b = \sum_{i=1}^{n_G^{(93)}} w_i^{(93)} I(Y_i^{(93)} = k, X_{i,sex}^{(93)} = x, E_i^{(93)} = 0). \tag{4.33}$$

The quantity s represents the prior sample size. We believe that the probability of reporting error is likely less than 20% and greater than .5%, because inevitably some respondents will accidentally select the wrong response. We interpret this statistically by requiring the 95% central credible interval of the prior to contain both .5% and 20%. To determine s for each (x, k), we create the following algorithm.

1. Set $s = 1$.

2. Let $a^* = s \times \frac{a}{a+b}$ and $b^* = s \times \frac{b}{a+b}$, with a and b defined as in (4.32) and (4.33).

3. Let F^{-1} be the inverse cumulative distribution function for a Beta distribution with parameters (a^*, b^*). Calculate $t_{025}(s) = F^{-1}(.025)$ and $t_{975}(s) = F^{-1}(.975)$.

4. If $t_{025}(s) \leqslant .005$ and $t_{975}(s) \geqslant .20$, then let $s = s + 1$. Repeat steps (2) - (4).

5. Otherwise, if $t_{025}(s) > .005$ or $t_{975}(s) < .20$, then s is too large. The final prior sample size is $s = s - 1$, the largest value of s for which both $t_{025}(s) \leqslant .005$ and $t_{975}(s) \geqslant .20$.

The prior sample size ranges from 6 to 18 for the error rate prior distributions for each gender and education level.

To construct priors for the reporting probabilities, we first compare the question wording of the 1990 Census and the 2010 ACS. In particular, we are concerned with the wording around professional degrees, because the 1993 NSCG data revealed that some respondents likely confused a job-related certification with a professional degree. Indeed, the 2010 ACS has clarified what is meant by professional degree. In the questionnaire, the check boxes for master's, professional, and doctorate degrees are all under a heading that states "After bachelor's degree." Additionally, the check box for professional degree states "Professional degree beyond a bachelor's degree (for example: MD, DDS, DVM, LLB, JD)." The "beyond a bachelor's degree" clause was not in the 1990 Census wording. Given this clarification in question wording, we expect that the probability of someone with no college degree reporting a professional degree will be lower in the 2010 ACS than in the 1990 Census. We suppose a priori that the improved wording reduces this case of reporting error by half. Otherwise, we expect the reporting probabilities in 2010 are likely to be similar to those in 1993, so our prior will be centered at the reporting probability estimate from the 1993 data.

More specifically, for each (x, k), our prior for $p_{x,k}(1, \ldots, d_Z)$, as defined in (4.10), is Dirichlet$\left(s \times \frac{a_{x,k}(1)}{\sum_{j=1}^{d_Z} a_{x,k}(j)}, \ldots, s \times \frac{a_{x,k}(d_Z)}{\sum_{j=1}^{d_Z} a_{x,k}(j)} \right)$ where

$$a_{x,k}(l) = \begin{cases} 0 & \text{if } k = l \\ \frac{1}{2} n_{x,k}(l) & \text{if } k = 5 \text{ (no college) and } l = 3 \text{ (prof. degree)} \\ n_{x,k}(l) & \text{otherwise} \end{cases} \tag{4.34}$$

$$\text{and} \quad n_{x,k}(l) = \sum_{i=1}^{n_G^{(93)}} w_i^{(93)} I(X_{i,sex}^{(93)} = x, Y_i^{(93)} = k, Z_i^{(93)} = l, E_i^{(93)} = 1). \tag{4.35}$$

The quantity s again represents the prior sample size. We want the prior to be fairly tight around the 1993 reporting probability estimates, which we interpret to mean that most of the prior weight is within 0.10 of the 1993 estimate. Specifically, the 95% central credible interval for each $p_{x,k}(l)$ for $k \neq l$ should include $\max\left(\frac{a_{x,k}(l)}{\sum_{j=1}^{d_Z} a_{x,k}(j)} - .1, 0\right)$ and $\min\left(\frac{a_{x,k}(l)}{\sum_{j=1}^{d_Z} a_{x,k}(j)} + .1, 1\right)$. We use the fact that the marginal distribution of each $p_{x,k}(l)$ for $k \neq l$ is Beta$(s \times \frac{a_{x,k}(1)}{\sum_{j=1}^{d_Z} a_{x,k}(j)}, s - s \times \frac{a_{x,k}(1)}{\sum_{j=1}^{d_Z} a_{x,k}(j)})$ (Albert and Denis, 2012). To determine s for each (x, k), we create the following algorithm.

1. Set $s = 1$.

2. For each $l \neq k$, let $a^*_{x,k}(l) = s \times \frac{a_{x,k}(1)}{\sum_{j=1}^{d_Z} a_{x,k}(j)}$, with $a_{x,k}(l)$ as defined above.

3. For each $l \neq k$, let F^{-1} be the inverse cumulative distribution function for a Beta distribution with parameters $(a^*_{x,k}(l), s - a^*_{x,k}(l))$. Calculate $t_{025,l}(s) = F^{-1}(.025)$ and $t_{975,l}(s) = F^{-1}(.975)$.

4. For each $l \neq k$, we check the following conditions:

   - $t_{025,l}(s) \leqslant \max\left(\frac{a_{x,k}(l)}{\sum_{j=1}^{d_Z} a_{x,k}(j)} - .1, 0\right)$

   - $t_{975,l}(s) \geqslant \min\left(\frac{a_{x,k}(l)}{\sum_{j=1}^{d_Z} a_{x,k}(j)} + .1, 1\right)$

   If these conditions are met for all $l \neq k$, then let $s = s + 1$. Repeat steps (2) - (4).

5. Otherwise, one of the credible intervals has become too narrow, meaning s is too large. The final prior sample size is $s = s - 1$, the largest value of s for which all of the marginal 95% central credible intervals of $p_{x,k}(1, \ldots, d_Z)$ contain the 1993 point estimate plus or minus 0.10.

The prior sample size ranges from 2 to 64 for males, and from 1 to 52 for females.

To evaluate the plausibility of this prior distribution, we impute Y using parameter draws from the prior. We sample 100 draws of β and $p_{x,k}(l)$ from the 1993 priors we just constructed. We also sample 100 draws of θ. Using these 100 prior parameter draws, we can create 100 imputations of Y. We can use the diagnostic check to see if the prior allows the model to match the distribution of Y given $\boldsymbol{X}$ as observed in the 2010 NSCG.

To use our diagnostic check, for each of the 100 completed datasets m we calculate $\hat{\pi}^m_{xk}$ for all 16 demographic combinations and 5 levels of NSCG education. We use MI combining rules to construct a 95% confidence interval for each $\boldsymbol{x}$ and k. We let $\hat{\pi}^G_{x,k} = \hat{T}^{(NSCG)}_{xk}/\hat{T}^{(ACS)}_{x,total}$ for $k = \{1:4\}$. For $k = 5$, $\hat{\pi}^G_{x5} = (\hat{T}^{(ACS)}_{x,total} - \sum_{j=1}^{4} \hat{T}^{(NSCG)}_{xj})/\hat{T}^{(ACS)}_{x,total}$. We calculate that 79 out of the 80 NSCG point estimates $\hat{\pi}^G_{xk}$ fall within the 95% confidence intervals from the prior-imputed ACS. Some of these intervals are quite large, but this diagnostic check assures us that the prior produces reasonable imputations of Y in the 2010 data.

The final model we consider for the sake of comparison is the CIA model. To impute Y under the CIA we use a Monte Carlo procedure where we sample θ^* and impute $(Y^*|\theta^*, \boldsymbol{X})$ in the ACS.

4.6.2 Results

As in our illustrative simulation, we initialize the missing NSCG-reported education Y to be equal to Z, the observed ACS-reported education. For each model specification, we save $M = 50$ completed ACS surveys. For each of the $M = 50$ completed datasets, we calculate the overall proportion of errors and the proportion of error by gender and NSCG education. Table 4.5 reports the average and 95% confidence interval for each of these error rates according to the model estimates. First we consider models 1 - 4, which are like those in Section 4.4 in that they do not have informative priors. Then we consider model 5, which uses the 1993 NSCG in the prior.

The CIA model estimates very high percentages of errors for all upper levels of

Table 4.5: Summary of results from different model specifications. Models 1-5 are run for 100000 MCMC iterations. We save $M = 50$ completed datasets under each model. For each dataset, we compute the estimated overall error rate, estimated error rate by gender and imputed Y, and associated variances using ratio estimation formulas (as in equations 4.25-4.28) that incorporate the ACS final survey weights. We combine results using standard MI combining rules; the table shows the mean and 95% confidence interval.

	estimated error rate: by group				estimated overall error rate
	Y=BA	Y=MA	Y=Prof.	Y=PhD	
CIA model					
Male	.37 (.36, .37)	.76 (.75, .76)	.91 (.91, .92)	.94 (.93, .95)	.57 (.55, .58)
Female	.35 (.35, .36)	.72 (.71, .72)	.95 (.94, .95)	.97 (.96, .97)	
Model 1					
Male	.05 (.04, .06)	.10 (.08, .11)	.18 (.15, .21)	.27 (.23, .31)	.17 (.16, .19)
Female	.05 (.05, .06)	.09 (.08, .10)	.18 (.15, .21)	.28 (.24, .32)	
Model 2					
Male	.05 (.04, .06)	.18 (.16, .21)	.27 (.18, .37)	.36 (.30, .42)	.20 (.18, .21)
Female	.05 (.05, .06)	.12 (.10, .14)	.26 (.20, .33)	.41 (.29, .53)	
Model 3					
Male	.05 (.04, .06)	.09 (.08, .11)	.17 (.14, .20)	.25 (.21, .30)	.17 (.16, .19)
Female	.05 (.05, .06)	.09 (.08, .10)	.17 (.14, .20)	.26 (.21, .31)	
Model 4					
Male	.05 (.04, .06)	.19 (.16, .23)	.36 (.26, .46)	.36 (.27, .45)	.22 (.20, .24)
Female	.09 (.08, .10)	.14 (.11, .17)	.52 (.44, .59)	.55 (.40, .70)	
Model 5					
Male	.07 (.06, .08)	.19 (.16, .22)	.23 (.14, .32)	.34 (.27, .41)	.22 (.20, .24)
Female	.09 (.08, .10)	.12 (.09, .15)	.50 (.43, .57)	.31 (.17, .46)	

education, indicating that the CIA is not plausible. Model 1 is the simplest of the other model specifications. Both model 2 and model 3 are generalizations of model 1; model 2 generalizes model 1 by allowing the probability of an error to depend on gender in addition to Y, and model 3 allows the probability of error to depend on race and Y. If the probability of error does not depend on gender in reality, then model 2 would produce similar results to model 1 in Table 4.5. However, we can see that model 2 estimates different error rates between males with master's degrees and females with master's degrees; males with master's degrees have a higher rate of reporting error in Z than their female counterparts. The analyst can similarly compare estimated error rates between races from model 3. It may be possible to allow the error model to depend on gender, race and Y, if there are enough free parameters.

Next we can explore the effect of changing the reporting model. Model 4 generalizes model 2 by allowing the reporting probabilities to vary by gender. In consultation with Dr. Sanders, we believe this model specification makes sense in context. If the reporting rates by gender were actually similar in reality, we would expect the two models to produce similar results. However, the estimated error rates are fairly different between model 2 and model 4; in particular the proportion of errors for female professionals in model 4 is estimated to be about double that of model 2. We can look at the reporting probabilities to see where the model is finding a difference, as in Table 4.6. Model 4 estimates some significant differences in reporting probabilities by gender. For example, males with bachelor's degrees who make a reporting error are estimated to report a master's degree with probability .96 whereas females with bachelor's degrees who make a reporting error are estimated to report a master's degree with probability .67 and a professional degree with probability .30. Other big differences exist for professional degree holders. Females with professional degrees who make a reporting error are most likely to report a bachelor's degree, whereas men with professional degrees who make a reporting error are most likely to report a master's degree or PhD.

We now turn to the results using the 1993 prior. Model 5 is different from models 1-4 in that we use informative prior distributions constructed from the 1993 NSCG data. We explore the influence of the prior by comparing the posterior of certain parameters under model 4 and model 5.

The prior influences the probability of reporting error more for females with PhDs

Table 4.6: Estimated mean and 95% confidence interval of reporting probabilities under model 2 and reporting probabilities by gender under model 4.

	Z=BA	Z = MA	Z = Prof.	Z = PhD
Y=BA				
m2	-	.95 (.87, 1.00)	.04 (.00, .11)	.01 (.00, .03)
m4 - Male	-	.96 (.90, 1.00)	.02 (.00, .07)	.02 (.00, .05)
m4 - Female	-	.67 (.58, .76)	.30 (.22, .38)	.03 (.00, .07)
Y=MA				
m2	.02 (.00, .06)	-	.51 (.43, .59)	.47 (.39, .55)
m4 - Male	.04 (.00, .11)	-	.57 (.48, .66)	.39 (.31, .47)
m4 - Female	.11 (.00, .25)	-	.39 (.26, .52)	.50 (.40, .61)
Y=Prof.				
m2	.05 (.00, .16)	.69 (.54, .83)	-	.26 (.14, .38)
m4 - Male	.02 (.00, .06)	.69 (.44, .94)	-	.29 (.04, .54)
m4 - Female	.91 (.79, 1.00)	.06 (.00, .16)	-	.04 (.00, .10)
Y= PhD				
m2	.01 (.00, .04)	.39 (.15, .63)	.60 (.36, .83)	-
m4 - Male	.01 (.00, .05)	.21 (.02, .39)	.78 (.60, .96)	-
m4 - Female	.10 (.00, .30)	.77 (.50, 1.00)	.13 (.00, .34)	-
Y=none				
m2	.95 (.95, .96)	.03 (.03, .04)	.01 (.01, .01)	.00 (.00, .00)
m4 - Male	.97 (.96, .97)	.03 (.02, .03)	.01 (.00, .01)	.00 (.00, .00)
m4 - Female	.96 (.95, .97)	.04 (.03, .05)	.00 (.00, .00)	.00 (.00, .00)

than for females with bachelor's degrees. This is probably because the data provide more information about bachelor degree holders. From the NSCG we estimate that of the females who reported at least a college degree in the ACS, roughly 58% of them have a bachelor's degree whereas only about 2% have a PhD. Figure 4.3 shows that the posterior probability of reporting error for females with bachelor's degrees is similar under both the 1993 prior and a uniform prior. Figure 4.4 shows that the posterior probability of reporting error for females with PhDs is noticeably different under the 1993 prior and a uniform prior.

Given the constraints provided by the two datasets, the parameter estimates are dependent on one another. We compare the reporting model probabilities of model

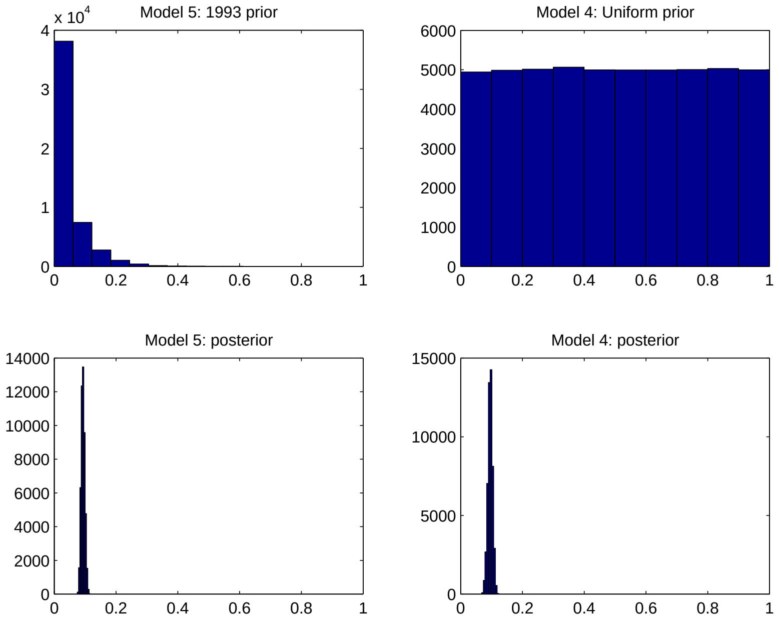


FIGURE 4.3: The top plots show 50000 draws of the probability of misreporting education for females with bachelor's degrees under the 1993 prior (model 5) and the uniform prior (model 4). The bottom plots show 50000 draws from the posterior under model 5 and model 4. The 1993 prior does not have much influence other than slightly tightening up the posterior credible interval.

4 and 5 to see how the prior influences these parameters. In particular we find that the prior has an influence on the reporting probabilities for individuals with graduate degrees. The posteriors under the 1993 prior and the uniform prior estimate a fairly similar probability of females reporting a bachelor's degree when the true education level is no college degree; see Figure 4.5.

For women with professional degrees who misreport education, the posterior probability of reporting a bachelor's degree under model 5 is slightly shifted from the prior, suggesting that the data do not provide much additional information about this reporting probability; see Figure 4.6. The posterior under the uniform prior of model 4 is quite different. Since there are few females with professional degrees, there is probably not strong information about this reporting probability. Without

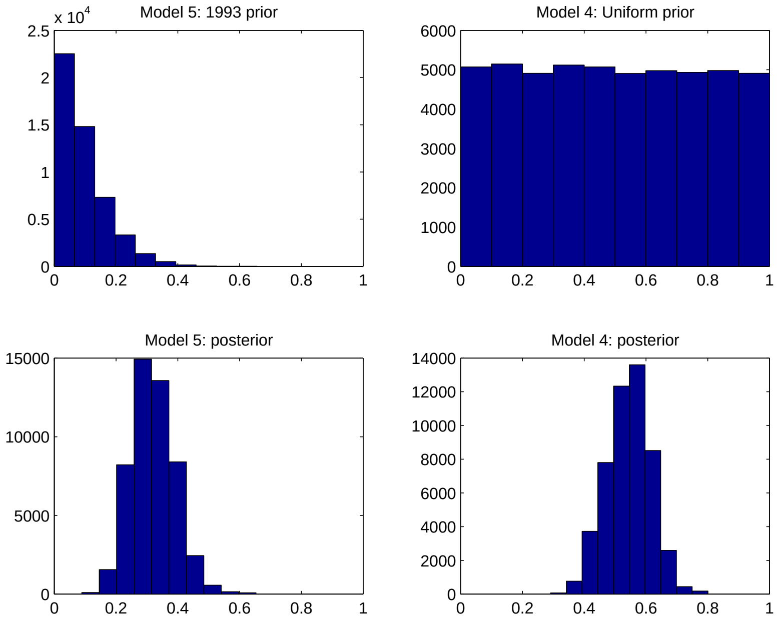


FIGURE 4.4: The top plots show 50000 draws of the probability of misreporting education for females with PhDs under the 1993 prior (model 5) and the uniform prior (model 4). The bottom plots show 50000 draws from the posterior under model 5 and model 4. The 1993 prior does have a clear influence on the posterior. This is due to the fact that there are fewer women with PhDs than women with bachelor's degrees.

informative prior information, the reporting probability is easily swayed by the other parameters in the model.

To decide between models 1, 2, 4, and 5, we can first use our diagnostic check to rule out any implausible models. To use our diagnostic check, for each completed datasets m we calculate $\hat{\pi}^m_{xk}$ for all 16 demographic combinations and 5 levels of NSCG education. We use MI combining rules to construct a 95% confidence interval for each $\boldsymbol{x}$ and k. We let $\hat{\pi}^G_{x,k} = \hat{T}^{(NSCG)}_{xk}/\hat{T}^{(ACS)}_{x,total}$ for $k = \{1 : 4\}$. For $k = 5$, $\hat{\pi}^G_{x5} = (\hat{T}^{(ACS)}_{x,total} - \sum_{j=1}^{4} \hat{T}^{(NSCG)}_{xj})/\hat{T}^{(ACS)}_{x,total}$. We calculate how many of the 80 NSCG point estimates $\hat{\pi}^G_{xk}$ fall within the 95% confidence interval from the ACS. The results are 73 out of 80 for model 1, 75 for model 2, 71 for model 3, 76 for model 4, 74 for

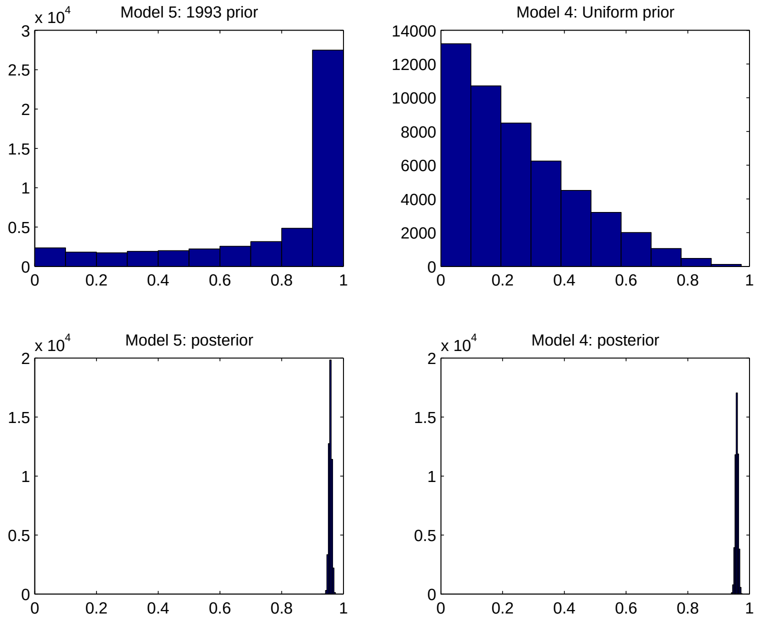


FIGURE 4.5: The top plots show 50000 draws of the probability of females reporting Z=bachelor's degree when Y=no college degree under the 1993 prior (model 5) and the uniform prior (model 4). The bottom plots show 50000 draws from the posterior under model 5 and model 4.

model 5, and 80 for the CIA model. The diagnostic does not suggest that any model is much worse than the others.

Considering the results reported above as well as the diagnostic, an analyst might choose model 5 even though it has more parameters to estimate compared to models 1 and 2. It seems plausible that the probability of misreporting education could depend on both gender and true education level, and it also seems plausible that the reported education level would also depend on gender and true education level. The model assumptions make sense in context, and the extra parameters allow the model to estimate differences between the genders and true education levels in both the error and reporting model. Additionally, the prior from the 1993 data seems to lead to more reasonable estimates in areas where there is not much information from

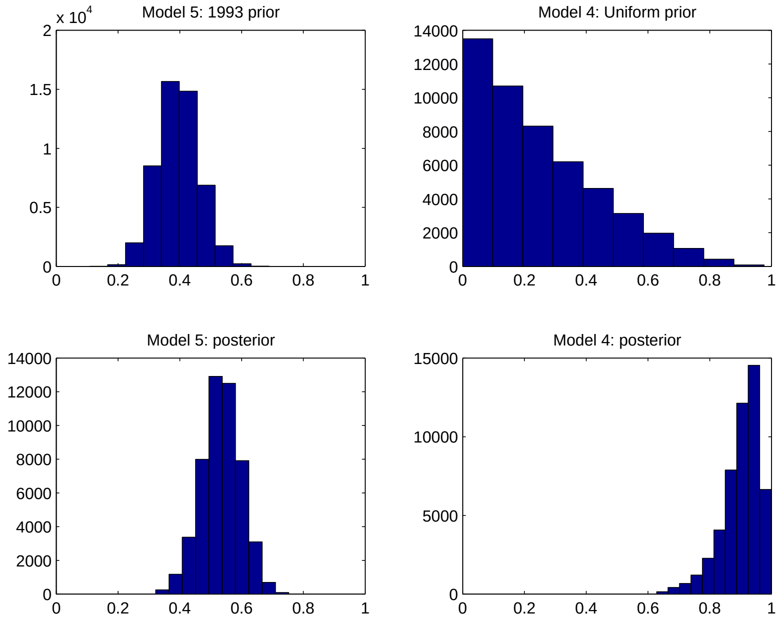


FIGURE 4.6: The top plots show 50000 draws of the probability of females reporting Z=bachelor's degree when Y=professional degree under the 1993 prior (model 5) and the uniform prior (model 4). The bottom plots show 50000 draws from the posterior under model 5 and model 4.

the data, e.g., for the more advanced degrees.

Alternatively, an analyst may determine that the estimated error rates by gender and Y seem too high in all of the models. In that case the analyst can specify different priors for the error probabilities and reporting probabilities. At one extreme, the analyst could fix the probability of an error to be 0 for individuals who truly have at least a college degree (Y = 1, 2, 3, or 4). The model would only estimate the reporting probabilities of individuals with no college degree ($Y = 5$). In other words, given the ACS-reported education Z, the true education Y would either be equal to Z or no college degree. Or the analyst could keep the model specification that makes the most sense and include stronger prior information for certain parameter estimates.

To answer our substantive questions, we calculate the average "wages or salary income past 12 months" (ACS variable "wagp" multiplied by "adjinc" to adjust for inflation) by gender and education level Y. We also calculate the total number of science and engineering degrees awarded to women (defined by the "sciengp" flag in the ACS) by each education level. We use the ACS final survey weights in our calculations. We combine results from the $M = 50$ completed datasets using MI combining rules. By comparing the results from the different model specifications, we can determine how sensitive our conclusions are to the model assumptions (Rubin, 1986).

Figure 4.7 shows the estimates of the total number of science and engineering degrees awarded to women. The estimates are fairly consistent across the models for bachelor's and master's degrees. Using the ACS-reported education would lead to higher estimates of the total number of females with science and engineering degrees at the professional and PhD level. At these degree levels, the conclusions seem fairly sensitive to the model specification. In particular, models 4 and 5 estimate a smaller number of science and engineering professional degrees being awarded to women compared to models 1-3. We use this discrepancy to explore why the models give different results.

The models give different results because they assign Y differently. When we estimate the total number of science and engineering degrees awarded to women for each level of education, the values for gender and science degree do not change. Only the distribution of science degrees awarded to women among the various true education levels can change with each model. The different error and reporting models imply a different distribution of Y conditional on Z and demographics $\boldsymbol{X}$.

We investigate how the different models impute women with degrees in science when they report a professional degree in the ACS. The plots in Figure 4.8 show the average number of women with science degrees imputed for different combinations of

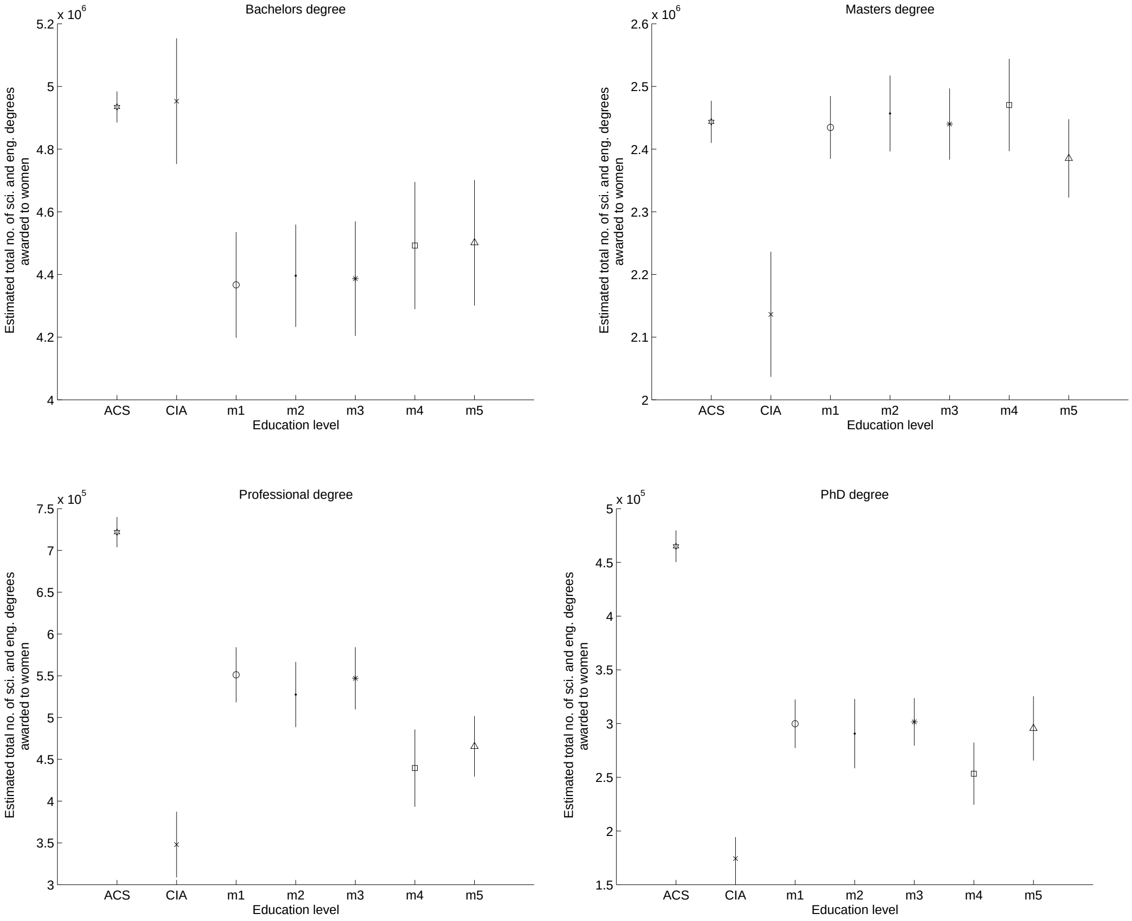


FIGURE 4.7: The estimated total number of science and engineering degrees awarded to women under each model. We plot the mean and 95% confidence intervals. Note the difference in scale for each degree category.

Y and Z when at least one of Y and Z is equal to professional degree. Compared to models 1-3, models 4 and 5 impute fewer true professional degrees for individuals who reported professional degrees. This finding lines up with our results from Table 4.5 where the estimated error rate for female professionals under models 4 and 5 was much larger than the other models. To meet the data constraints it seems that models 4 and 5 put more weight on the combinations (Y=Prof, Z=BA) and (Y=BA, Z=Prof) for women than the other models.

Our other question of interest was about how income varies with education level.

Figure 4.9 shows the estimates of average income. Looking at the estimated average income for professional degrees, using the ACS-reported education gives the highest estimate of average income, and the CIA model gives the lowest estimate of average income. There is quite a bit of difference across the other model estimates. The models vary a lot in how they impute professional degrees in Y, leading to the differences we are seeing in number of female science degrees and average incomes.

Figure 4.10 shows the estimates of average male income and average female income. The prior in model 5 has a clear effect on average income for PhDs; the gap between average male and female income is more pronounced in model 4 than in model 5.

4.7 Conclusion

The statistical framework presented here can help researchers assess how sensitive their conclusions are to different assumptions about reporting error. In the ACS/NSCG example, the conclusions are somewhat sensitive to the different model assumptions. In the future, the NSCG organizers may wish to consider releasing the ACS-reported education along with the NSCG responses at least for some records. This would allow researchers to observe the true reporting error process and use that information to impute a more accurate education response in the ACS. This would benefit anyone who uses the ACS-reported education in their research.

When there are a large number of variables with many levels, it may become difficult to specify the error and reporting models. A future direction may be to incorporate the error and reporting model structure within the DPMPM model; for example, Si et al. (2015a) allow some of the multinomial distributions within each latent class to depend on an indicator variable for attrition.

It may also become difficult to specify the conditional model $(Y|\boldsymbol{X})$ when there are many covariates. It may be possible to apply some ideas from Schifeling and

Reiter (2016) to jointly model $(\boldsymbol{X}, Y)$ even when the NSCG has complex survey design. For example, we could draw T^*_{xk} for each covariate group d and education level k as in Section 4.5. However, rather than calculating θ^*_{xk}, we could create a synthetic margin of T^*_{xk} observations for each d and k. Schifeling and Reiter (2016) explain how to incorporate a synthetic margin of observations in the DPMPM model.

4.8 Posterior computation

The joint model for the data and parameters can be written as follows. Let n be the total sample size of both datasets. For ease of notation, we assume the reporting probabilities do not depend on $\boldsymbol{X}$, so that $p_{x,k}(l) = p_k(l)$ for all covariate combinations x.

$$
\begin{aligned}
&p\left(\boldsymbol{X}_{1:p}, Y, Z, E | \Theta, \beta, \{p_k(l)\}\right) \times p\left(\theta\right) p\left(\beta\right) p\left(\{p_k(l)\}\right) \\
&= \left(\prod_{i=1}^{n} p\left(\boldsymbol{X}_i, Y_i | \theta\right) \cdot p\left(E_i | \boldsymbol{X}_i, Y_i, \beta\right) \cdot p\left(Z_i | \boldsymbol{X}_i, Y_i, E_i, \{p_k(l)\}\right)\right) \\
&\quad \times p\left(\theta\right) p\left(\beta\right) p\left(\{p_k(l)\}\right) \\
&= \left(\prod_{i=1}^{n} p(\boldsymbol{X}_i, Y_i | \theta)\right) \left(\prod_{\substack{i=1 \\ Y_i \neq d_Z + 1}}^{n} \left(\Phi(M_i^T \beta)\right)^{I(E_i=1)} \left(1 - \Phi(M_i^T \beta)\right)^{I(E_i=0)}\right) \\
&\quad \times \left(\prod_{\substack{i=1 \\ E_i=1}}^{n} p_{Y_i}(Z_i)\right) \times p\left(\theta\right) p\left(\beta\right) p\left(\{p_k(l)\}\right).
\end{aligned}
$$

- Sample from posterior of θ, parameters in DPMPM:

$$
p(\theta | ...) \propto \left(\prod_{i=1}^{n} p(\boldsymbol{X}_i, Y_i | \theta)\right) p(\theta).
$$

For details, see Appendix A.

- Sample from posterior of β, coefficients in error model:

$$p(\beta|...)\propto\left(\prod_{\substack{i=1\\Y_i\neq d_Z+1}}^{n}\left(\Phi(M_i^T\beta)\right)^{I(E_i=1)}\left(1-\Phi(M_i^T\beta)\right)^{I(E_i=0)}\right)p(\beta).$$

One option is to introduce latent variables as in Albert and Chib (1993). In our models, we specified an error model with all interaction terms, so that $\boldsymbol{M}$ essentially partitioned the sample into mutually exclusive subsets. If $\boldsymbol{M}[i,j]=1$, then unit i is in subset j. For each j we sample $\Phi(\beta_j)\sim\text{Beta}(n_{1,j}+b_{1,j},n_{0,j}+b_{0,j})$, where

$$n_{1,j}=\sum_{\substack{i=1,\\\boldsymbol{M}[i,j]=1}}^{n}(E_i=1)$$

and

$$n_{0,j}=\sum_{\substack{i=1,\\\boldsymbol{M}[i,j]=1}}^{n}(E_i=0).$$

The prior for $\Phi(\beta_j)$ is $\text{Beta}(b_{1,j},b_{0,j})$. If $b_{1,j}=b_{0,j}=1$, this is a uniform prior for the probability of reporting error for group j.

- For $k\in\{1:d_Y\}$, sample from posterior of $\{p_k(\cdot)\}$, probabilities in reporting model:

$$\begin{aligned}p(\{p_k(\cdot)\}|...)&=\left(\prod_{l=1}^{d_Z}(p_k(l))^{\sum_{i=1}^n I(Y_i=k,Z_i=l,E_i=1)}\right)p\left(\{p_k(\cdot)\}\right)\\&=\text{Dirichlet}\left(\alpha_k(1),\ldots,\alpha_k(d_Z)\right),\end{aligned}$$

where $\alpha_k(l)=\sum_{i=1}^n I(Y_i=k,Z_i=l,E_i=1)+a_{kl}$ for $k\neq l$. For a uniform prior, each $a_{kl}=1$ for $k\neq l$. For $k=l$, $\alpha_k(k)=0$ so that $p_k(k)=0$.

- Fill in Z in the NSCG. Z_i can take value l for $l \in \{1 : d_Z\}$. We write $M_i(k)$ instead of M_i to specify that the model matrix for the probit model can depend on Y.

$$Pr(Z_i = l|Y_i = k, M_i(k), \{p_k(l)\}, ...) = \begin{cases} 1 - \Phi(M_i(k)^T\beta) & \text{if } l = k \text{ and } k \in \{1 : d_Z\} \\ \Phi(M_i(k)^T\beta) \cdot p_k(l) & \text{if } l \neq k \text{ and } k \in \{1 : d_Z\} \\ p_{d_Z+1}(l) & \text{if } k = d_Z + 1. \end{cases} \tag{4.36}$$

- Fill in Y in the Census. Y_i can take value k for $k \in \{1 : d_Z + 1\}$. First note:

$$\begin{aligned} Pr(Y_i = k|\boldsymbol{X}_i, Z_i = l, ...) &\propto Pr(Y_i = k, \boldsymbol{X}_i, Z_i = l) \\ &\propto Pr(Z_i = l|Y_i = k, M_i(k), \{p_k(l)\}, \ldots) \\ &\quad \times Pr(Y_i = k, \boldsymbol{X}_i). \end{aligned}$$

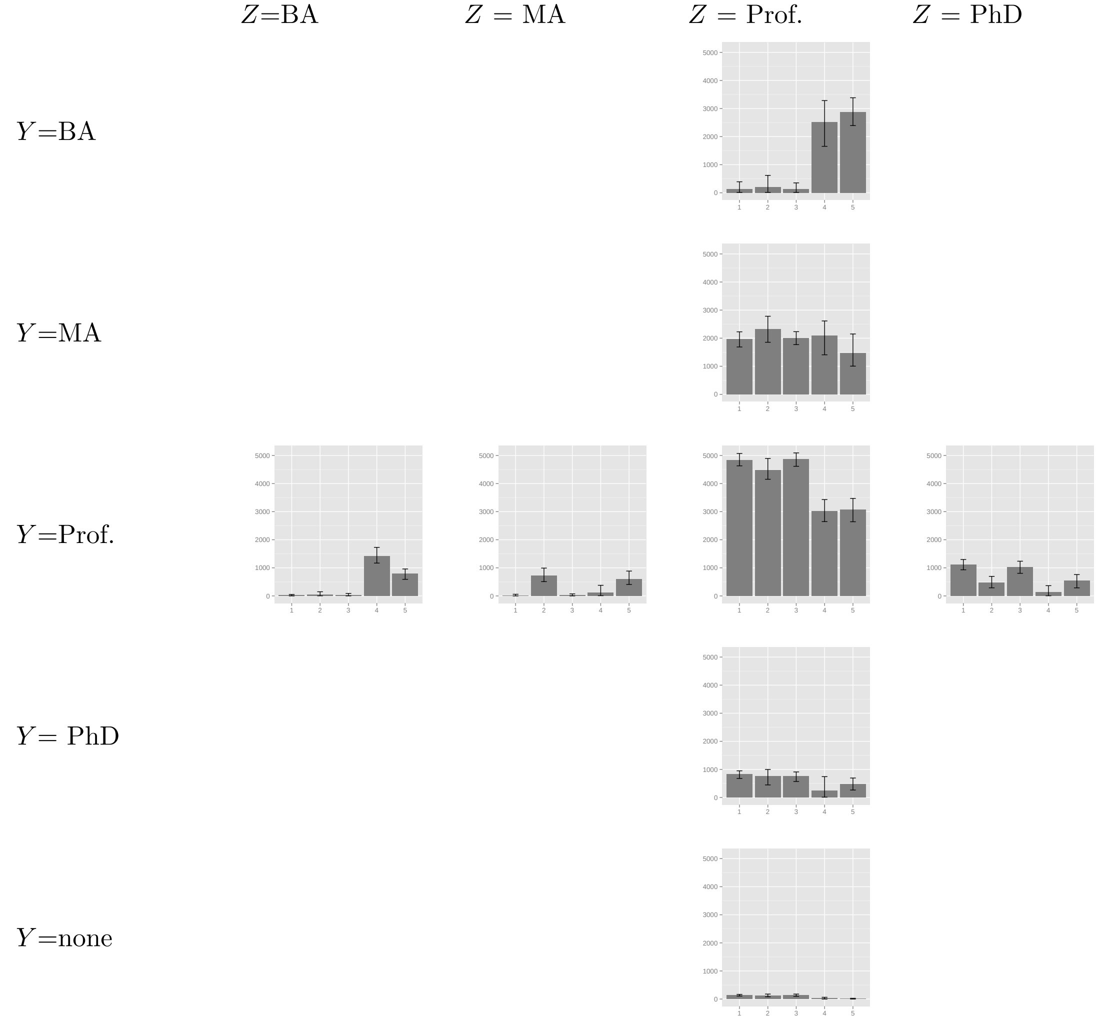
  From the true data model we get $Pr(Y_i = k, \boldsymbol{X}_i)$. For example, if using the DPMPM model we get:

$$Pr(Y_i = k, \boldsymbol{X}_i) = \sum_{h=1}^{H*} \pi_h \left(\phi_{h,NSCG,k} \prod_{j=1}^{p} \phi_{hjX_{ij}} \right).$$

  $Pr(Z_i = l|Y_i = k, M_i(k), \{p_k(l)\}, \ldots)$ is as in Equation 4.36.

- Compute $E_i = I(Y_i \neq Z_i)$.

Figure 4.8: Number of females with science degrees imputed in each combination of Y and Z, where at least one of Y and Z is a professional degree. Bar plots the average over $M = 50$ multiple imputations, error bars show 2.5% and 97.5% quantiles.



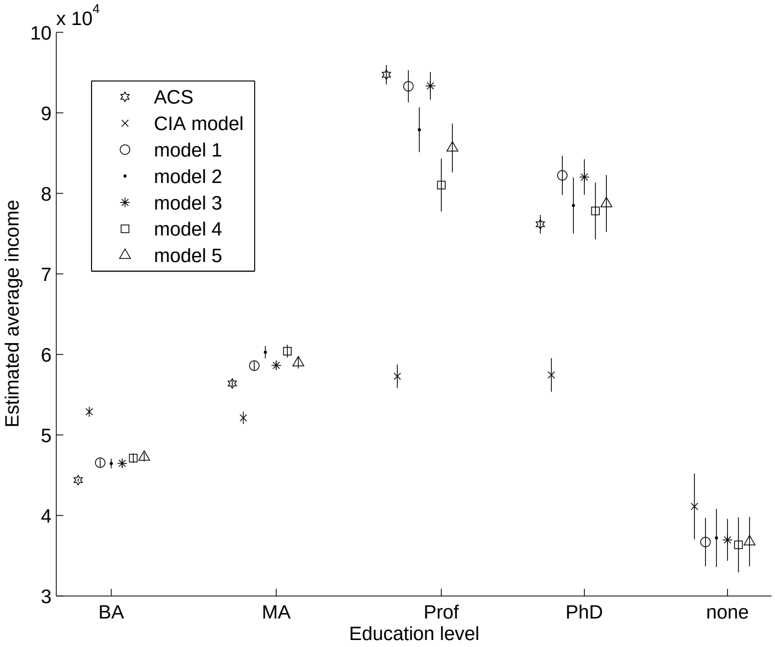


FIGURE 4.9: For each model, within each education level k and completed dataset m, we calculate the average income. Note that some individuals have income of 0; these individuals are still included in the calculations. We combine the estimates using MI combining rules, and we plot the mean and 95% confidence intervals. For the ACS estimate we calculate the average income within each ACS-reported education level.

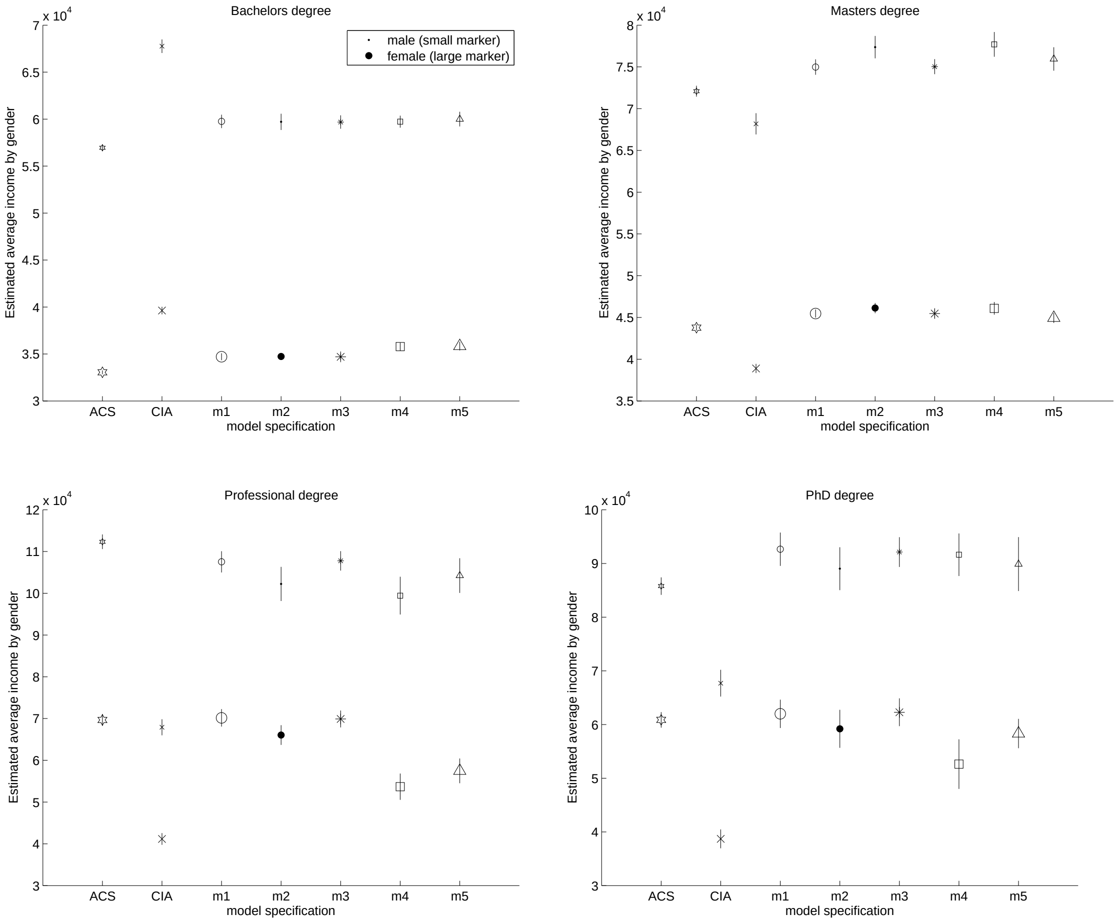


FIGURE 4.10: For each model, within each education level k and completed dataset m, we calculate the average income by gender. Note that some individuals have income of 0; these individuals are still included in our calculations. We combine the estimates using MI combining rules, and we plot the mean and 95% confidence intervals. For the ACS estimate we calculate the average income within each ACS-reported education level.

5

Conclusion and Future Directions

This thesis presented Bayesian approaches to combining information from multiple surveys. Chapter 2 presented a method to account for information from a refreshment sample when modeling nonignorable unit nonresponse and attrition. Chapter 3 presented an approach to incorporate marginal prior information in latent class models for categorical survey data. Chapter 4 developed a reporting error model that combines information from a gold standard survey to impute true responses in survey with reporting error.

In reality, many surveys simultaneously have missing data from unit nonresponse or attrition, complex sampling designs, and reporting error. Additionally, an analyst can collect S sources of information, which can be in the form of survey data, marginal prior information, and additional surveys such as refreshment samples. The ideal model would incorporate the information from all S data sources when imputing missing data due to nonresponse or reporting error. We describe a future research direction of how the ideas in this thesis can be fused into one survey analysis method.

Let $\boldsymbol{X}$ be the collection of demographic variables in any of the S data sources, let $\boldsymbol{Y}$ be the collection of all outcome variables, and let $\boldsymbol{Z}$ be the collection of all reported responses that may be subject to reporting error. If a particular variable is not recorded in data source s for $s \in \{1 : S\}$, then it is coded as missing data. This approach is similar to how Gelman et al. (1998a) proposed combining information from multiple surveys when not all surveys asked the same questions.

We can imagine extending the ideas of Chapter 3 to account for various types of complex sampling, so that each data source s can be augmented with a margin that balances the sampling design. The method of Dong et al. (2014a), which combined information from multiple complex surveys by generating synthetic populations for each survey, may also be useful here.

It may be possible to adapt the ideas of Chapter 3 to impute missing covariates from nonignorable unit nonresponse, an issue we dealt with in Chapter 2. Pfeffermann and Sikov (2011) present a model that incorporates known population totals to impute missing covariates.

Once each data source s has been augmented with a margin to account for sampling design, we specify a joint model for all S data sources as in Gelman et al. (1998a). It will be important to determine which parameters are survey-specific and which are universal for all surveys. For example, the joint distribution for $(\boldsymbol{X}, \boldsymbol{Y})$ is likely the same for all surveys if we assume they are all samples from the same population. However, the different surveys may have different reporting error mechanisms, so the reporting error model specifications and parameters may be survey-specific.

A flexible joint model for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ would allow an analyst to combine information from all S data sources, thus incorporating all relevant information about sampling designs, informative marginal or joint prior distributions, and plausible reporting error mechanisms. The analyst could use the model parameters to make inference on the distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ as well as save multiple imputations of any data source s.

Appendix A

Posterior computation of DPMPM model

Chapters 3 and 4 make use of the DPMPM model. We use a Gibbs sampler to estimate the posterior distributions of the unknown quantities $(z, V, \pi, \alpha, \phi)$. The full conditionals are similar to those in Si and Reiter (2013).

The steps of the Gibbs sampler are as follows.

1. To update z_i for $i = 1, \ldots, n$, sample from a categorical distribution with

$$p(z_i = h | X_{i1}, \ldots X_{ip}, \pi, \phi) = \frac{\pi_h \prod_{j=1}^{p} \phi_{hjX_{ij}}}{\sum_{k=1}^{H*} \pi_k \prod_{j=1}^{p} \phi_{kjX_{ij}}}. \tag{A.1}$$

2. To update V_h for $h = 1, \ldots, H^* - 1$, we sample from

$$p(V_h | \alpha, z) = \text{Beta}\left(n_h + 1, \alpha + \sum_{k=h+1}^{H*} n_k\right) \tag{A.2}$$

where $n_h = \sum_{i=1}^{n+n_A} 1(z_i = h)$. Set $V_{H*} = 1$. Then, $\pi_h = V_h \prod_{g<h}(1 - V_g)$ for $h = 1, \ldots, H^*$.

3. To update α, sample from

$$p(\alpha|V_1, \ldots, V_{H^*-1}) = \text{Gamma}\left(H^* + a_\alpha - 1, b_\alpha - \log\left(\pi_{H^*}\right)\right). \tag{A.3}$$

4. To update ϕ_{hj} for variable j and $h = 1, \ldots, H^*$, sample from

$$p(\phi_{hj}|X_{obs}, z) = \text{Dirichlet}\left(1 + \sum_{\substack{i=1 \\ z_i=h}}^{n} 1(X_{ij} = 1), \ldots, 1 + \sum_{\substack{i=1 \\ z_i=h}}^{n} 1(X_{ij} = d_j)\right). \tag{A.4}$$

Bibliography

Agresti, A. (2013), *Categorical Data Analysis*, Wiley-Interscience, Hoboken, NJ, 3rd edn.

Aitchison, J. and Shen, S. M. (1980), "Logistic-normal distributions: Some properties and uses," *Biometrika*, 67, 261–272.

Albert, I. and Denis, J.-B. (2012), "Dirichlet and multinomial distributions: properties and uses in Jags," Tech. Rep. 2012-5, INRA France.

Albert, J. H. and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88, 669–679.

Baltagi, B. H. and Song, S. H. (2006), "Unbalanced panel data: A survey," *Statistical Papers*, 47, 493–523.

Bartels, L. M. (1999), "Panel effects in the American national election studies," *Political Analysis*, 8, 1–20.

Bayarri, M. J. and Berger, J. O. (1998), "Quantifying Surprise in the Data and Model Verification," *Bayesian Statistics*, 6, 53 – 82.

Bayarri, M. J. and Berger, J. O. (2000), "P Values for Composite Null Models," *Journal of the American Statistical Association*, 95, 1127–1142.

Behr, A., Bellgardt, E., and Rendtel, U. (2005), "Extent and determinants of panel attrition in the European Community Household Panel," *European Sociological Review*, 21, 489–512.

Bhattacharya, D. (2008), "Inference in panel data models under attrition caused by unobservables," *Journal of Econometrics*, 144, 430–446.

Black, D., Sanders, S., and Taylor, L. (2003), "Measurement of Higher Education in the Census and Current Population Survey," *Journal of the American Statistical Association*, 98, 545–554.

Black, D., Haviland, A., , Sanders, S., and Taylor, L. (2006), "Why do minority men earn less? A study of wage differentials among the highly educated," *The Review of Economics and Statistics*, 88, 300–313.

Black, D. A., Haviland, A. M., , Sanders, S. G., and Taylor, L. J. (2008), "Gender Wage Disparities among the Highly Educated," *Journal of Human Resources*, 43, 630–659.

Brick, J. M. and Kalton, G. (1996), "Handling missing data in survey research," *Statistical Methods in Medical Research*, 5, 215 – 238.

Buckley, J. and Schneider, M. (2006), "Are Charter school parents more satisfied with schools? Evidence from Washington, DC," *Peabody Journal of Education*, 81, 57–78.

Burgette, L. F. and Reiter, J. P. (2010), "Multiple imputation via sequential regression trees," *American Journal of Epidemiology*, 172, 1070–1076.

Carrig, M. M., Manrique-Vallier, D., Ranby, K., Reiter, J. P., and Hoyle, R. (2015), "A multiple imputation-based method for the retrospective harmonization of data sets," *Multivariate Behavioral Research*, 50, 383–397.

Cohen, M. P. (1997), "The Bayesian bootstrap and multiple imputation for unequal probability sample designs," *Proceedings of the Survey Research Methods Section*, American Statistical Association, 635 – 638.

Curran, P. J. and Hussong, A. M. (2009), "Integrative Data Analysis: The Simultaneous Analysis of Multiple Data Sets," *Psychological Methods*, 14, 81–100.

Das, M., Toepoel, V., and van Soest, A. (2011), "Nonparametric Tests of Panel Conditioning and Attrition Bias in Panel Surveys," *Sociological Methods and Research*, 40, 32–56.

Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., and Zheng, S. (2013), "Handling attrition in longitudinal studies: The case for refreshment samples," *Statistical Science*, 28, 238 – 256.

D'Errico, J. (2013), "nearestSPD: Finding the nearest positive definite matrix," MATLAB Central File Exchange, Retrieved February 11, 2016.

Dobra, A., Tebaldi, C., and West, M. (2006), "Data augmentation in multi-way contingency tables with fixed marginal totals," *Journal of Statistical Planning and Inference*, 136, 355–372.

Dominici, F., Parmigiani, G., Reckhow, K. H., and Wolpert, R. L. (1997), "Combining Information from Related Regressions," *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 313 – 332.

Dong, Q., Elliott, M. R., and Raghunathan, T. E. (2014a), "Combining information from multiple complex surveys," *Survey Methodology*, 40, 347–354.

Dong, Q., Elliott, M. R., and Raghunathan, T. E. (2014b), "A nonparametric method to generate synthetic populations to adjust for complex sampling design features," *Survey Methodology*, 40, 29–46.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006), *Statistical Matching: Theory and Practice*, Wiley, Hoboken, NJ.

D'Orazio, M., Di Zio, M., and Scanu, M. (2010), "Old and new approaches in statistical matching when samples are drawn with complex survey designs," *Proceedings of the 45th Scientific Meeting of the Italian Statistical Society*.

D'Orazio, M., Di Zio, M., and Scanu, M. (2012), "Statistical matching of data from complex sample surveys," *Proceedings of the European Conference on Quality in Official Statistics*.

Dunson, D. B. and Bhattacharya, A. (2011), "Nonparametric Bayes Regression and Classification Through Mixtures of Product Kernels," in *Bayesian Statistics 9, Proceedings of Ninth Valencia International Conference on Bayesian Statistics*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford University Press.

Dunson, D. B. and Xing, C. (2009), "Nonparametric Bayes modeling of multivariate categorical data," *Journal of the American Statistical Association*, 104, 1042–1051.

Fesco, R. S., Frase, M. J., and Kannankutty, N. (2012), "Using the American Community Survey as the Sampling Frame for the National Survey of College Graduates," Working Paper NCSES 12-201, National Science Foundation, National Center for Science and Engineering Statistics, Arlington, VA.

Finamore, J. (2013), *National Survey of College Graduates: About The Survey*, National Center for Science and Engineering Statistics.

Fosdick, B. K., DeYoreo, M., and Reiter, J. P. (2015), "Categorical Data Fusion Using Auxiliary Information," *arXiv:1506.05886 [stat.ME]*.

Gebregziabher, M. and DeSantis, S. M. (2010), "Latent class based multiple imputation approach for missing categorical data," *Journal of Statistical Planning and Inference*, 140, 3252 – 3262.

Gelman, A. (2007), "Struggles with Survey Weighting and Regression Modeling," *Statistical Science*, 22, 153–164.

Gelman, A. and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457–472.

Gelman, A., King, G., and Liu, C. (1998a), “Not Asked and Not Answered: Multiple Imputation for Multiple Surveys,” *Journal of the American Statistical Association*, 93, 846 – 857.

Gelman, A., King, G., and Liu, C. (1998b), “Not Asked and Not Answered: Multiple Imputation for Multiple Surveys: Rejoinder,” *Journal of the American Statistical Association*, 93, 869 – 874.

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

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., and Meulders, M. (2005), “Multiple imputation for model checking: Completed-Data plots with missing and latent data,” *Biometrics*, 61, 74–85.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008), “A weakly informative default prior distribution for logistic and other regression models,” *The Annals of Applied Statistics*, 2, 1360–1383.

Goodman, L. A. (1974), “Exploratory latent structure analysis using both identifiable and unidentifiable models,” *Biometrika*, 61, 215–231.

Goyder, J. (1987), *The silent minority: Nonrespondents on sample surveys*, Polity Press, Cambridge.

Greenland, S. (2007), “Prior data for non-normal priors,” *Statistics in Medicine*, 26, 3578–3590.

Groves, R. M., Singer, E., and Corning, A. (2000), “Leverage-saliency theory of survey participation: description and an illustration,” *The Public Opinion Quarterly*, 64, 299–308.

Groves, R. M., Dillman, D., Eltinge, J. L., and Little, R. J. (2002), *Survey nonresponse*, Wiley, New York.

Groves, R. M., Presser, S., and Dipko, S. (2004), “The role of topic interest in survey participation decisions,” *Public Opinion Quarterly*, 68, 2–31.

Guo, Y. and Little, R. J. (2011), “Regression analysis with covariates that have heteroscedastic measurement error,” *Statistics in Medicine*, 30, 2278 – 2294.

He, Y. and Zaslavksy, A. M. (2009), “Combining information from cancer registry and medical records data to improve analyses of adjuvant cancer therapies,” *Biometrics*, 65, 946–952.

He, Y., Zaslavsky, A. M., and Landrum, M. B. (2010), “Multiple imputation in a large-scale complex survey: A guide,” *Statistical Methods in Medical Research*, 19, 653–670.

He, Y., Landrum, M. B., and Zaslavksy, A. M. (2014), "Combining information from two data sources with misreporting and incompleteness to assess hospice-use among cancer patients: a multiple imputation appraoch," *Statistics in Medicine*, 33, 3710–3724.

Hillygus, D. S. (2005), "Campaign effects and the dynamics of turnout intention in election 2000," *Journal of Politics*, 67, 50–68.

Hinkins, S., Oh, H. L., and Scheuren, F. (1994), "Inverse Sampling Design Algorithms," in *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Hirano, K., Imbens, G. W., Ridder, G., and Rubin, D. B. (1998), "Combining panel data sets with attrition and refreshment samples," Tech. Rep. 230, National Bureau of Economic Research.

Hirano, K., Imbens, G., Ridder, G., and Rubin, D. (2001), "Combining panel data sets with attrition and refreshment samples," *Econometrica*, 69, 1645–1659.

Hoff, P. D. (2009), *A First Course in Bayesian Statistical Methods*, Springer, New York.

Hogan, J. W. and Daniels, M. J. (2008), *Missing Data in Longitudinal Studies*, Chapman and Hall, Boca Raton.

Horvitz, D. G. and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663 – 685.

Hu, J. (2015), "Dirichlet Process Mixture Models for Nested Categorical Data," Ph.D. thesis, Department of Statistical Science, Duke University.

Imbens, G. W. and Pizer, W. A. (2000), "The Analysis of Randomized Experiments with Missing Data," Resources for the Future, Discussion Paper 00-19.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American Statistical Association*, 96, pp. 161–173.

Jackson, C. H., Best, N. G., and Richardson, S. (2009), "Bayesian graphical models for regression on multiple data sets with different variables," *Biostatistics*, 10, 335 – 351.

Jain, S. and Neal, R. M. (2004), "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, 13, 158 – 182.

Johndrow, J., Cron, A., and Dunson, D. B. (2014), "Bayesian tensor factorizations for massive web networks," in *ISBA World Meeting 2014 in Cancun, Mexico*.

Kalli, M., Griffin, J. E., and Walker, S. G. (2009), "Slice sampling mixture models," *Statistics and Computing*, 21, 93–105.

Kamakura, W. A. and Wedel, M. (1997), "Statistical data fusion for cross-tabulation," *Journal of Marketing Research*, 34, 485–498.

Kessler, D. C., Hoff, P. D., and Dunson, D. B. (2014), "Marginally specified priors for non-parametric Bayesian estimation," *Journal of the Royal Statistical Society: Series B*.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P., and Wang, Q. (2015), "Simultaneous edit-imputation for continuous microdata," *Journal of the American Statistical Association*, 110, 987 – 999.

Kunihama, T. and Dunson, D. B. (2013), "Bayesian modeling of temporal dependence in large sparse contingency tables," *Journal of the American Statistical Association*, 108, 1324–1338.

Kunihama, T., Herring, A. H., Halpern, C. T., and Dunson, D. B. (2014), "Nonparametric Bayes modeling with sample survey weights," *arXiv:1409.5914v2*.

Lazar, R., Meeden, G., and Nelson, D. (2008), "A noninformative Bayesian approach to finite population sampling using auxiliary variables," *Survey Methodology*, 34, 51–64.

Little, R. and Rubin, D. (2002), *Statistical Analysis with Missing Data*, Wiley-Interscience, Hoboken, N.J.

Little, R. J. (2003), "The Bayesian Approach to Sample Survey Inference," in *Analysis of Survey Data*, eds. R. L. Chambers and C. J. Skinner, chap. 4, pp. 49 – 57, Wiley, Hoboken, NJ.

Little, R. J. A. and Zheng, H. (2006), "The Bayesian Approach to the Analysis of Finite Population Surveys," in *Bayesian Statistics 8*, pp. 282 – 302.

Liu, C. (2004), "Robit Regression: A Simple Robust Alternative to Logistic and Probit Regression," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, eds. A. Gelman and X.-L. Meng, John Wiley and Sons, Ltd.

Lohr, S. L. (2010), *Sampling: Design and Analysis*, Brooks/Cole, Boston, MA, 2nd edn.

Lynn, P. (2013), “Alternative sequential mixed-mode designs: effects on attrition rates, attrition bias, and costs.” *Journal of Survey Statistics and Methodology*, 1, 183–205.

Manrique-Vallier, D. and Reiter, J. P. (2014a), “Bayesian estimation of discrete multivariate latent structure models with structural zeros,” *Journal of Computational and Graphical Statistics*, 23, 1061 – 1079.

Manrique-Vallier, D. and Reiter, J. P. (2014b), “Bayesian multiple imputation for large-scale categorical data with structural zeros,” *Survey Methodology*, 40, 125–134.

Manrique-Vallier, D. and Reiter, J. P. (2015), “Bayesian simultaneous edit and imputation for multivariate categorical data,” `http://mypage.iu.edu/~dmanriqu/papers/LCM_Zeros_EdImp.pdf`.

Meng, X. (1994), “Posterior predictive p-values,” *Annals of Statistics*, 22, 1142–1160.

Molenberghs, G., Goetghebeur, E. J. T., Lipsitz, S. R., and Kenward, M. G. (1999), “Nonrandom Missingness in Categorical Data: Strengths and Limitations,” *The American Statistician*, 53, 110–118.

Molenberghs, G., Kenward, M. G., and Goetghebeur, E. (2001), “Sensitivity Analysis for Incomplete Contingency Tables: The Slovenian Plebiscite Case,” *Journal of the Royal Statistical Society, Series C*, 50, 15 – 29.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M. G. (2008), “Every missingness not at random model has a missingness at random counterpart with equal fit,” *Journal of the Royal Statistical Society, Series B*, 70, 371–388.

Moriarity, C. and Scheuren, F. (2001), “Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure,” *Journal of Official Statistics*, 17, 407 – 422.

National Research Council (2008), *Using the American Community Survey for the National Science Foundation’s Science and Engineering Workforce Statistics Programs*, Panel on Assessing the Benefits of the American Community Survey for the NSF Division of Science Resources Statistics, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, The National Academies Press, Washington, DC.

National Science Foundation (1993), “National Survey of College Graduates, 1993,” http://doi.org/10.3886/ICPSR06880.v1, ICPSR06880-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2014-10-02.

Nevo, A. (2003), “Using weights to adjust for sample selection when auxiliary information is available,” *Journal of Business and Economic Statistics*, 21, 43–52.

Olsen, R. J. (2005), "The problem of respondent attrition: Survey methodology is key," *Monthly Labor Review*, 128, 63–71.

Papaspiliopoulos, O. (2008), "A note on posterior sampling from Dirichlet mixture models (Technical Report)," *Centre for Research in Statistical Methodology*.

Pepe, M. S. (1992), "Inference using surrogate outcome data and a validation sample," *Biometrika*, 79, 355 – 365.

Peytchev, A. (2013), "Consequences of survey nonresponse," *The ANNALS of the American Academy of Political and Social Science*, 645, 88–111.

Pfeffermann, D. (1993), "The Role of Sampling Weights When Modeling Survey Data," *International Statistical Review*, 61, 317 – 337.

Pfeffermann, D. (2011), "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?" *Survey Methodology*, 37, 115–136.

Pfeffermann, D. and Sikov, A. (2011), "Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information," *Journal of Official Statistics*, 27, 181 – 209.

Prior, M. (2010), "You've either got it or you don't? The stability of political interest over the life cycle," *The Journal of Politics*, 72, 747–766.

Raghunathan, T. E. (2006), "Combining information from multiple surveys for assessing health disparities," *Allgemeines Statistisches Archiv*, 90, 515–526.

Raghunathan, T. E. and Grizzle, J. E. (1995), "A Split Questionnaire Survey Design," *Journal of the American Statistical Association*, 90, 54 – 63.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003), "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, 19, 1–16.

Rao, J. N. K., Scott, A. J., and Benhin, E. (2003), "Undoing Complex Survey Data Structures: Some Theory and Applications of Inverse Sampling," *Survey Methodology*, 29, 107 – 128.

Rassler, S. (2002), *Statistical Matching*, Springer, New York.

Reiter, J. P. (2002), "Satisfying disclosure restrictions with synthetic data sets," *Journal of Official Statistics*, 18, 531–544.

Reiter, J. P. (2012), "Bayesian finite population imputation for data fusion," *Statistica Sinica*, 22, 795 – 811.

Rodriguez, G. (2007), "Lecture Notes on Generalized Linear Models," .

Rosenbaum, P. R. (2010), *Observational Studies*, Springer, New York.

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

Rubin, D. B. (1981), "The Bayesian Bootstrap," *The Annals of Statistics*, 9, 130–134.

Rubin, D. B. (1986), "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations," *Journal of Business & Economic Statistics*, 4, 87–94.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Rubin, D. B., Stern, H. S., and Vehovar, V. (1995), "Handling "Don't Know" Survey Responses: The Case of the Slovenian Plebiscite," *Journal of the American Statistical Association*, 90, 822 – 828.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall, New York.

Schenker, N. and Raghunathan, T. E. (2007), "Combining information from multiple surveys to enhance estimation of measures of health," *Statistics in Medicine*, 26, 1802–1811.

Schenker, N., Raghunathan, T. E., and Bondarenko, I. (2010), "Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey," *Statistics in Medicine*, 29, 533–545.

Schifeling, T. A. and Reiter, J. P. (2016), "Incorporating Marginal Prior Information in Latent Class Models," *Bayesian Analysis*, 11, 499–518.

Schifeling, T. A., Cheng, C., Reiter, J. P., and Hillygus, D. S. (2015), "Accounting for nonignorable unit nonresponse and attrition in panel studies with refreshment samples," *Journal of Survey Statistics and Methodology*, 3, 265 – 295.

Schwartz, S. L., Li, F., and Reiter, J. P. (2012), "Sensitivity analysis for unmeasured confounding in principal stratification," *Statistics in Medicine*, 31, 949–962.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Si, Y. and Reiter, J. P. (2013), "Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys," *Journal of Educational and Behavioral Statistics*, 38, 499–521.

Si, Y., Reiter, J. P., and Hillygus, D. S. (2015a), "Bayesian Latent Pattern Mixture Models for Handling Attrition in Panel Studies with Refreshment Samples," *arXiv:1509.02124v1*.

Si, Y., Pillai, N. S., and Gelman, A. (2015b), "Bayesian Nonparametric Weighted Sampling Inference," *Bayesian Analysis*, 10, 605–625.

Si, Y., Reiter, J. P., and Hillygus, D. S. (2015c), "Semi-parametric Selection Models for Potentially Non-ignorable Attrition in Panel Studies with Refreshment Samples," *Political Analysis*, 23, 92–112.

Singer, E. (2006), "Introduction Nonresponse Bias in Household Surveys," *Public Opinion Quarterly*, 70, 637–645.

Smith, T. W., Marsden, P., Hout, M., and Kim, J. (2015), "General Social Surveys, 1972-2012: Cumulative Codebook," National Data Program for the Social Sciences Series, No. 22.

Tarmast, G. (2001), "Multivariate Log-Normal Distribution," in *International Statistical Institute: Seoul 53rd Session*.

Traugott, M. W. and Tucker, C. (1984), "Strategies for predicting whether a citizen will vote and estimation of electoral outcomes," *Public Opinion Quarterly*, 48, 330–343.

Vermunt, J. K., Van Ginkel, J. R., Van Der Ark, L. A., and Sijtsma, K. (2008), "Multiple imputation of incomplete categorical data using latent class analysis," *Sociological Methodology*, 38, 369–397.

Wade, S., Mongelluzzo, S., and Petrone, S. (2011), "An Enriched Conjugate Prior for Bayesian Non-parametric Inference," *Bayesian Analysis*, 6, 359 – 385.

Walker, S. G. (2007), "Sampling the Dirichlet mixture model with slices," *Communications in Statistics - Simulation and Computation*, 36, 45–54.

Western, B. (2002), "The impact of incarceration on wage mobility and inequality," *American Sociological Review*, 67, 526–546.

Yucel, R. M. and Zaslavsky, A. M. (2005), "Imputation of binary treatment variables with measurement error in administrative data," *Journal of the American Statistical Association*, 100, 1123–1132.

Zhang, G., Schenker, N., Parker, J. D., and Liao, D. (2013), "Identifying implausible gestational ages in preterm babies with Bayesian mixture models," *Statistics in Medicine*, 32, 2097–2113.

Zhou, H., Elliott, M. R., and Raghunathan, T. E. (2015), "A Two-Step Semiparametric Method to Accommodate Sampling Weights in Multiple Imputation," *Biometrics*.

Zhou, J., Bhattacharya, A., Herring, A. H., and Dunson, D. B. (2014), "Bayesian factorizations of big sparse tensors," *Journal of the American Statistical Association*, p. to appear.

Zhou, X. and Reiter, J. P. (2010), "A note on Bayesian inference after multiple imputation," *The American Statistician*, 64, 159–163.

Biography

Tracy Anne (Boswell) Schifeling was born on February 17, 1989 in Mount Holly, N.J. She received a B.A. in Mathematics from the University of Chicago in June 2010, an *en route* M.S. in Statistical Science from Duke University in May 2014, and a Ph.D. in Statistical Science from Duke University in 2016.

She was a SAMSI graduate fellow with the Computational Methods in Social Sciences program during the 2013-14 academic year. In summer 2014, she was a Data Science for Social Good fellow in Chicago.

Her published papers include Schifeling et al. (2015) and Schifeling and Reiter (2016).