University of California
Los Angeles

Sequential Bayesian Regression for Multiple Imputation and Conditional Editing

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Public Health

by

Robin Angela Jeffries

2013

Abstract of the Dissertation

Sequential Bayesian Regression for Multiple Imputation and Conditional Editing

by

Robin Angela Jeffries

Doctor of Public Health

University of California, Los Angeles, 2013

Professor Robert E. Weiss, Chair

Analysts faced with errors in data apply editing rules to fix erroneous data. These edits are deterministically assigned and edits may not be correct in all cases. This dissertation presents a unified method to multiply impute missing data and multiply edit erroneous data using a sequence of Bayesian regression models. The techniques used to multiply edit erroneous data are an exact parallel for multiple imputation used to correct missing data. The models presented allow for different data types subject to several error mechanisms.

This method is called Sequential Bayesian Regression for Multiple Imputation and Conditional Editing (SyBRMICE) and creates multiple fully imputed and edited data sets. Desired analyses are performed on each complete and consistently edited and imputed data set individually. Results from these analyses are combined using the same combining rules used in multiple imputation. The resulting parameter estimates and intervals will then correctly account for the errors incurred in both the data editing and imputation processes.

Development of SyBRMICE was motivated by data from Project Connect (PC). Project Connect was an 8 year longitudinal intervention study aiming to reduce teen pregnancy and STD rates in select middle and high schools in the Los Angeles area. Survey data was collected annually to measure the effectiveness of the interventions. A paper survey was administered to the students as a group in the classroom, and student responses have both missing and erroneous data.

The Project Connect survey was administered annually for five years. A subset of students participated in multiple years resulting in repeated answers to the same question by the same student. Data errors found in the PC survey data can be categorized as belonging to one of several error types. If a variable such as gender that should remain constant over time is observed to differ across surveys, this variable then is said to have an inconsistent longitudinal response. If a variable, such as age or ever having sexual intercourse, that should increase monotonically over time is observed to have a non-monotonic reporting pattern, this variable is then said to have an inconsistent monotonic longitudinal response. Lastly if the responses to two or more related variables give conflicting information, these variables are said to have an inconsistent multiple response.

Models to stochastically edit each of the three types of erroneous data are presented. The inconsistent repeated measures, inconsistent monotone longitudinal, and inconsistent multivariate models are developed separately and then combined as steps in an example of the larger unifying SyBRMICE procedure. The examples demonstrate the flexibility and customizability of the SyBRMICE procedure. Results from an analysis performed on the multiple complete and consistent data sets generated by the SyBRMICE procedure are compared to results from the same analysis performed on a single deterministically-edited, complete-case data set.

The dissertation of Robin Angela Jeffries is approved.

Peter R. Kerndt

Thomas R. Belin

Abdelmonem A. Afifi

Robert E. Weiss, Committee Chair

University of California, Los Angeles

2013

To my nieces, nephews, and new little sister;

Never think that you are unable to achieve something because of where you come from or where you are at in life. Whatever you have been through and will go through, you are strong and will come out stronger. Find your goals, your passions and live them. Don't settle for mediocrity, and never be bored. To use a quote from a beloved novel,

Your life is yours alone, rise up and live it. (Goodkind, 2001)

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to acknowledge my dissertation committee for supporting me in this research endeavour, with a special thanks to my chair Dr. Weiss. His continual patience, availability, encouragement and quirky humor have been instrumental to my success. I would like to thank Dr. Belin for always managing to find financial support when it was needed, allowing me to achieve this goal. I also express my gratitude to the entire Adolescent and School Health unit at the LACDPH DHSP. Especially Dr. Christine De Rosa who has been unbelievably supportive during my entire program. I extend my thanks and love to my mom and dad for raising me with a sense of responsibility and drive to accomplish things. Most importantly I want to thank my husband for putting up with me, supporting me, and especially marrying me, during this tremendously difficult time in my life.

Vita

1993-1996	High School, Chico Senior High School, Chico, CA
2005	B.S. (Biology, Statistics), CSUC, Chico, California.
2007	M.S. (Biostatistics), UCLA, Los Angeles, California.
2005–2008, 2012	Teaching Assistant, UCLA Biostatistics, Los Angeles, California
2006	Project Assistant, UCLA Center for Public Health and Disasters, Los Angeles, California
2006–2008	Biostatistics Graduate Student Researcher, Behavioral Epidemiology Research Group, Los Angeles, California
2007–2011	Data Manager/Analyst, Project Connect, Health Research Association, Los Angeles, California
2011–present	Statistician/Database Administrator, Keeping It Real LAC, Los Angeles County Department of Public Health, Division of HIV and STD Programs, and Institute for Health Promotion and Disease Prevention Research, University of Southern California, Los Angeles, California

Publications

DeRosa CJ, Jeffries RA, Afifi AA, Cumberland WG, Chung EQ, Kerndt PK, Ethier KA, Martinez E, Loya RV, Dittus PJ. (2012). "Improving the implementation of a condom availability program in urban high schools." *Journal of Adolescent Health*, 51, 5, 572–579. DOI: http://dx.doi.org/10.1016/j.jadohealth.2012.03.010

Jyrala A, Weiss RE, Jeffries RA, Kay GL. (2012). "Is hypothyroidism overlooked in cardiac surgery patients?" *Open Journal of Thoracic Surgery*, 2, 29–35. DOI: 10.4236/ojts.2012.22009

Gorbach PM, Weiss RE, Fuchs E, Jeffries RA, Hezerah M, Brown S, Voskanian A, Robbie E, Anton P, Cranston RD. (2012). "The slippery slope: lubricant use and rectal sexually trans-

mitted infections: A newly identified risk." *Sexually Transmitted Diseases*, 39, 1, 59–64. DOI: 10.1097/OLQ.0b013e318235502b

Finley DS, Shuch B, Said JW, Galliano G, Jeffries RA, Afifi AA, Castor B, Sadaat A, Kabbinavar FF, Belldegrun AS, Pantuck AJ. (2011). "The chromophobe tumor grading system is the preferred grading scheme for chromophobe renal cell carcinoma." *Journal of Urology*, 186, 6, 2168–2174. DOI: 10.1016/j.juro.2011.07.068

Gorbach PM, Weiss RE, Jeffries RA, Javanbakht M, Drumright LN, Daar ES, Little SJ. (2011). "Behaviors of recently HIV-infected men who have sex with men in the year post-diagnosis: effects of drug use and partner types." *Journal of Acquired Immune Deficiency Syndromes*, 56, 2, 176–182. DOI: 10.1097/QAI.0b013e3181ff9750

Jyrala A, Weiss RE, Jeffries RA, Kay GL. (2010). "Effect of mild renal dysfunction on presentation, short-and long-term outcomes on on-pump cardiac surgery patients." *Interactive CardioVascular Thoracic Surgery* DOI:10.1510/icvts.2009.231068

Micheletti R, Fishbein G, Fishbein M, Singer E, Weiss RE, Jeffries RA, Currier JS. (2009). "Coronary atherosclerotic lesions in human immunodeficiency virus-infected patients: a histopathologic study." *Cardiovascular Pathology*, 18, 28–36.

CHAPTER 1

Introduction to Project Connect

Project Connect's official title was "Integrated, Multi-Level Interventions to Improve Adolescent Health through the Prevention of Sexually Transmitted Diseases, Including HIV, and Teen Pregnancy." The primary aim of the study was to reduce teen pregnancy and STD rates in adolescents by implementing multiple structural intervention plans. Structural interventions are designed to intervene in the environment the people live in, not on the person themselves. The study population consisted of students who attended select middle and high schools in areas of Los Angeles with higher than the 2004 national average of teen births and rates of Chlamydia.

1.1 Aims

Project Connect (PC) had a two-armed study design. Twelve high schools and 14 of their corresponding feeder middle schools were equally split between the intervention and control conditions. The baseline year was the 2005-06 academic year, with all interventions beginning fall of the 2006-07 academic year. The research team designed and implemented structural interventions for the parents, schools, community, and local health care providers.

The parent intervention included mailing pamphlets and DVDs home to the parents of students attending intervention schools. These materials provided information for the parents on how to keep track of their adolescent children after school, how to talk with them about sensitive subjects including sex and pregnancy, and ways they could be more involved in their child's life.

The school intervention assisted the implementation and management of the campus based

Condom Availability Program (CAP). These CAPs are mandated by the school district, but implemented at various levels of compliance (DeRosa et al., 2012, Rafferty and Radosh, 1997). School level interventions also included health teacher training events designed to boost the effectiveness of teachers in teaching health education topics.

The community intervention connected community services to students by creating a resource guide containing details about after school activities in the community. These guides were given to school staff for distribution to students. Project staff also facilitated bringing the Los Angeles County Health Department's mobile testing unit to intervention high-school campuses without a school-based health center in an effort to increase STD testing and awareness of services.

Provider level interventions included meetings with school administration and nurses to discuss the need for reproductive health care services, and annual Link-over-Lunch meetings to connect nurses and providers in conversations on better ways to serve youth. PC staff also created a provider referral guide for the school nurses. This guide listed adolescent friendly reproductive health care providers in the community to whom they could refer students for follow-up reproductive health care.

1.2 Evaluation

To evaluate the effectiveness of the Project Connect (PC) intervention a self-report survey was conducted annually during the spring semesters from 2005 to 2009 in the participating schools. Students were asked to report on their attitudes, feelings and behaviors regarding sensitive matters such as sexual behaviors and drug use. This dissertation uses data collected from the first four years only.

Each Fall semester entire classrooms within both control and intervention schools were selected using a random cluster sampling design. Every student within the selected classes was invited to participate in the evaluation survey for that year. Students in the 6^{th} , 8^{th} , and 10^{th} grade classes selected at baseline, and the 8^{th} grade classes selected during the second year, were specifically asked to participate for the duration of the study as longitudinal

	2005	2006	2007	2008
Non-responder	51.9%	49.5%	41.4%	46.2%
Refused consent	7.2%	6.7%	6.2%	4.9%
Consented	40.9%	43.8%	52.4%	49.0%

Table 1.1: Percent of invited students who did not return their consent form, who actively refused consent and who consented to participate in the Project Connect survey.

cohort participants.

After the classrooms were selected, student roster data was obtained from the school for all students enrolled in those classes. This roster data consisted of the student's class schedule and personal information including name, home address, gender, date of birth, and ethnicity. This information was used to mail study information to their homes, to track students in the longitudinal cohort across years and in some of the procedures discussed in Section 3.5.1.

PC staff administered the self-report survey to students in the selected classrooms during Spring of each school year. The paper survey was taken by the student during a single class period. Since less than half of the invited students consented to participate, not all students present at the time of data collection took the survey. Occasionally non-participating students were disruptive to or attempted to collaborate with the participating students.

1.2.1 Sample Size

Active parental informed consent was required from every student. Students over 18 could sign or refuse the consent form on their own. Table 1.1 gives consent rates during the first four years. Parental consent forms were given out to every student in the selected classrooms. Those that never returned their consent form were considered non-responders. Few, 7.2%, 6.7%, 6.2% and 4.9% refused consent in 2005, 2006, 2007, and 2008, respectively. These students, or their parents, specifically chose not to participate in this study. All students had the option to refuse to assent to the survey on the day of administration. Table 1.2

	2005	2006	2007	2008
Cross-Sectional (Single Survey)	6686	4461	4518	5188
Longitudinal (Repeated Measures)	3792	4630	4464	2834
Total	10478	9091	8982	8022

Table 1.2: Sample size by analysis cohort over time.

presents the number of surveys collected during the first four years. Multiple surveys were collected on 5,998 students resulting in a total of 15,720 observations that can be used in repeated measures analyses. An additional 20,853 surveys were collected from cross-sectionally selected students for a total of 36,573 surveys on 26,851 students. Ages of the students ranged from 11 to 18 years old, with a mean of 14.9 and standard deviation of 2.1.

1.2.2 Survey Instrument

The Project Connect survey is approximately 40 pages with 18 sections as listed in Table 1.3. This survey was based on the Youth Risk Behavior Surveillance (YRBS) (Centers for Disease Control and Prevention, 2010), a national survey of adolescents. Reliability studies for the YRBS survey have been done by Brener et al. (1995), Davey et al. (2001) and Troped et al. (2007). Some individual questions and some sections are only asked of high school students. The students marked their responses by filling in bubbles corresponding to the chosen response option, or wrote text in the space provided as applicable. These forms were processed by a Teleform scanning system that converted the image to an electronic data file.

Section	Number of Questions	Section Name
A	(24)	About You
B	(12)	School Activities
C	(12)	Your School
D	(2)	Health Care
E	(4)	School-Based Health Center (SBHC) (HS only)
F	(4)	Your School Nurse
G	(6)	Condom Availability Program (CAP) (HS only)
H	(13)	Reproductive Health Care (HS only)
I	(16)	Sexual Activity
J	(6)	Plans about Sexual Activity
K	(6)	Parental Monitoring
L	(8)	Friends and Dating
M	(6)	Family Communication, Part 1
N	(8)	Family Communication, Part 2
O	(6)	Adults in Your Life
P	(6)	Other People's Opinions
Q	(5)	Your Neighborhood
R	(11)	Your Health

Table 1.3: Sections of the Project Connect survey.

1.3 Missing Data

When a survey question is not answered, it is considered *missing data*. The teleform scanning software included a visual verification process to ensure that all marks on the paper survey are correctly translated into a database. Therefore, missing data that occurs is due to the student not answering the question.

Missing data can be problematic for an analyst. The more data that is missing the lower the power will be to draw conclusions from the data. Imputation is a technique used to fill in the missing values with data generated by a chosen process, so that the values are no longer missing. Section 3.1 provides an introduction to missing data imputation and discusses some imputation processes. Imputation is a now commonplace method of handling missing data, but should be done in a thoughtful manner to account for the fact that the imputation process adds data that was not there before.

1.4 Inconsistent Data

Self report surveys rely on the participant responding accurately. There is no consequence to the student for not reporting accurately, nor for reporting inconsistently across years. *Inconsistent* data is when the reported responses to multiple questions, or the same question asked multiple times, provide conflicting information. When the reported responses do not represent truth, that is, when the observed values do not equal the true values, these values are *erroneous*. Inconsistent data is *erroneous* data, but not all data that are erroneous are inconsistent. Inconsistent data can be identified using data cleaning procedures I discuss later. Other terminology used for inconsistent data is that a variable is subject to a *data error, reporting error* or *mis-reporting*.

Erroneous data can arise for a number of reasons that may not be distinguishable. Questions could be written in a confusing fashion or the participants may fail to recall past events accurately. The participant may not take the survey seriously and may fill in arbitrary responses. In addition to the more, perhaps benign, reasons for mis-reporting, cognitive

psychology research indicates that the participant weighs the true response to the question against their feelings, and what they perceive to be the socially desirable response (Cannell et al., 1981). Depending on factors such as, but not limited to age, religion, upbringing and perceptions of peer beliefs, certain behaviors such as wearing a seat belt, drug use, and sexual activity could be considered either socially desirable or undesirable. Participants may decide to not answer a question or to answer in a manner that is less than truthful. The underlying reasons for why people respond or choose not to respond is unknown. The combination of missing and erroneous responses observed in the PC data set provide the motivation for this dissertation.

Standard data cleaning processes for categorical variables include examining the variables on a univariate basis. This is done to identify data entry errors such as observing a value of 2 for a variable that should only contain 0 and 1 as response options. However, survey questions do not exist in a vacuum. Many questions are associated with, or even directly related to other questions. In this dissertation I frequently use the term *response pattern*. This refers to the combination of responses to more than one question considered jointly. For example a response pattern for a student who received an "A" grade in English class and a "B" grade in Math would look like (A,B). Similarly a response pattern for a person who reported being Male on two surveys would look like (M,M).

This is where the multivariate categorical nature of the Project Connect survey can make finding errors in the data tricky. Often there is no indication that a response is in error when the variable is examined univariately. Only when the response pattern to multiple questions is examined is the error visible.

I consider three types of inconsistent responses; responses to a question that change over time when they shouldn't change, responses that change over time in an impossible way, and responses to multiple questions that give conflicting or inconsistent information.

1.4.1 Inconsistent Repeated Measurements

Inconsistent Repeated Measures (IRM) occur when the responses given to a single repeated question asked over time are inconsistent, or otherwise provide conflicting information. I distinguish between two different cases for IRM, those that are time fixed, and those with a structured trend. Time fixed variables are ones where there is a strong prior belief or physical restriction that there is only one true underlying fixed value that does not change over time. Structured trend variables are ones that are allowed to change over time, but only in a specified way such as monotone increasing or decreasing. Inconsistencies observed in variables with a structured trend I call *Inconsistent Monotone Longitudinal* responses, or IML.

Examples of Project Connect survey variables that are considered time fixed include gender, birth date, number of elementary schools attended, and the students'/mothers'/fathers' birthplace. Time fixed variables tend to be demographics and often are used as subject level predictors and as stratifying variables. An example of an IRM would be when a student reports being male for the first year, then female for the second and third year surveyed (M,F,F). The challenge in addressing these errors comes in trying to assess which value is in error. For a student who responds male for two years and female for two years (M,M,F,F), how do you determine what the true gender of the student is?

Examples of Project Connect survey variables that are allowed to change over time include age, grade, number of high schools attended, lifetime usage of health care facilities, ever having intercourse, number of lifetime partners, and lifetime use of alcohol/tobacco or other drugs. These variables are allowed to remain constant, or change over time but only in a monotone increasing manner.

1.4.2 Inconsistent Monotone Longitudinal Responses

Some variables such as age or grade are expected to increase linearly over time. Binary variables such as ever smoking a cigarette or ever having sexual intercourse are a 1-step step-function where the true response starts off No (coded as 0), and stays No until it

changes to Yes (coded as 1) and remains Yes for the remainder of the study. An example of an Inconsistent Monotone Longitudinal response error would be if a student said No the first year, Yes the second, and then switched back to No when surveyed a third time (0,1,0).

A procedure to correct these inconsistent repeated measures should be able to estimate the true underlying value and it should also properly represent the uncertainty in that estimate. Two detailed examples of inconsistent repeated measures are provided in Chapter 4, and an inconsistent monotone longitudinal response is modeled in Chapter 7.

1.4.3 Inconsistent Multivariate Responses

Inconsistent multivariate (IMV) responses occur when responses to multiple variables that should be consistent, aren't, or otherwise give conflicting information. Bivariate examples include when biologically implausible combinations of height (3' 0") and weight (399 lbs) occur, or responding No to ever smoked but Yes to smoking in the past month.

Consider the questions that define a student as sexually experienced (`I1:` *"Have you ever had sexual intercourse"*) and sexually active (`I6:` *"Have you had sexual intercourse in the past 3 months"*). Both are Yes (1) / No (0) binary response questions. Examining these variables univariately would only verify that no data entry errors, such as a response of 2, had occurred. An inconsistent bivariate error is only discovered by looking at the combination of responses between the experience question and active question.

A student who marks "No" they have never had sexual intercourse, and then subsequently marks "Yes" they have had sex in the past 3 months provides an inconsistent response pattern. This response combination is physically impossible, at least one of the two responses to the variables in question is incorrect. This is an example of why there is a need to correct the responses prior to analysis. If the student really is not sexually experienced, then they would be incorrectly included in an analysis of active students. If they really are sexually active, then they would be incorrectly excluded from an analysis of experienced students.

Many analyses (for example DeRosa et al., 2012, Ethier et al., 2011 and Habel et al., 2010) using the Project Connect data set are done on selected subsets of students such as students

who reported knowing about the School Based Health Center, students who reported ever having sex, or those who reported having had sex in the past 3 months. If variables with reporting errors are used in prediction models, it will result in a reduction in accuracy.

Data editing is any change made to the data from its original input form. Data are edited to correct inconsistent or erroneous data by changing the responses from values that are considered to be incorrect to values considered to be correct. A common method for editing inconsistent data uses subject matter theory to decide what was "meant" to be reported and changing one or more values accordingly.

As the number of interrelated variables increases so does the potential for erroneous response patterns. Section 3.5.2 provides a detailed example of many conflicting answers in conjunction with missing values. Information contained in related variables can be valuable in any editing or imputation process to deal with inconsistent or missing data.

The rest of this dissertation is arranged as follows: Chapter 2 gives a description of the Bayesian paradigm and defines terms and algorithms. Chapter 3 provides a review of the imputation and editing processes, variable notation, and discusses current imputation and editing procedures in the Project Connect data set. Models and subsequent editing and imputation procedures are provided for two IRM examples in Chapter 4, an IML example in Chapter 5 and an example of an IMV between two binary variables example in Chapter 6.

Chapter 7 wraps it all together in a unified model to multiply edit and multiply impute multivariate missing and inconsistent data. Chapter 8 covers assumptions, limitations and extensions of this work. Sections from the Project Connect codebook that provide full text versions of questions discussed in this paper can be found in Appendix A.

CHAPTER 2

Bayesian Analysis and Simulation Methods

This chapter provides a brief review of Bayesian inference and Markov chain Monte Carlo methods to define notation and algorithms used in this dissertation. Full discussions of these topics can be found in textbooks by Gelman et al. (2004), Robert and Casella (2005), and Albert (2009). The purpose of using these algorithms is to make inference regarding a parameter θ after observing data y. This is done by simulating a sample from the posterior $p(\theta|y)$ and calculating summary statistics such as the mean, variance, and 95% intervals. These summaries then are used to make inference regarding $\theta|y$.

2.1 Bayes Theorem

Bayesian inference treats all parameters as random. Both the data y and parameter θ have distributions. To make inference about the parameter θ given the data y you create a full joint probability model $p(\theta, y)$. The model should be consistent with knowledge about the underlying scientific problem and the data collection process (Gelman et al., 2004). Using the rules of conditional probability and solving for $p(\theta|y)$ gives

$$\begin{aligned} p(\theta|y)p(y) &= p(\theta, y) = p(\theta)p(y|\theta), \\ \frac{p(\theta|y)p(y)}{p(y)} &= \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \\ p(\theta|y) &= \frac{p(\theta)p(y|\theta)}{p(y)}, \end{aligned} \tag{2.1}$$

where $p(\theta|y)$ is the posterior distribution of the parameter θ conditional on the observed data y, and $p(y)$ is the marginal distribution of the data. The parameter θ has a prior distribution $p(\theta)$, and $p(y|\theta)$ is the sampling distribution of y conditional on θ. The last line of (2.1)

is known as Bayes' Theorem. Since the denominator $p(y)$ is a constant this is commonly written

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{2.2}$$

This states that the posterior distribution of the parameter conditional on the observed data is proportional to the product of the prior distribution and the sampling distribution.

When analyses are performed on data y_i from subjects $i = 1, \ldots, n$ the data are represented as a vector $\mathbf{y} = (y_1, \ldots, y_n)'$. Since the desired inference is on the parameter $\theta|\mathbf{y}$, the data distribution $p(\mathbf{y}|\theta)$ can be viewed as a function of θ instead of $\mathbf{y}$

$$\begin{aligned} p(\mathbf{y}|\theta) &= \prod_{i=1}^{n} p(y_i|\theta) \\ &= L(\theta|\mathbf{y}), \end{aligned} \tag{2.3}$$

where $L(\theta|\mathbf{y})$ is the likelihood of θ given y. Equation (2.3) assumes independence between subjects. Bayes' theorem (2.2) can then be written as the posterior is proportional to the prior times the likelihood

$$p(\theta|\mathbf{y}) \propto p(\theta)L(\mathbf{y}|\theta). \tag{2.4}$$

When the statistical model is complicated or high dimensional, algebraically calculating summaries of the posterior density $p(\theta|y)$ can be difficult or numerically intractable. When the posterior cannot be directly sampled from easily or at all, iterative sampling methods such as Markov chain Monte Carlo (MCMC) methods can be used to generate a sample from $p(\theta|y)$.

2.2 Markov Chain Monte Carlo Sampling

Markov chain Monte Carlo (MCMC) methods are iterative sampling algorithms which produce a sequence of values $\mathbf{S} = \{\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(L)}\}$ that hold specific properties (Robert and Casella, 2005). One of these properties is that the sequence $\mathbf{S}$, or *chain*, converges to a stationary distribution f. This convergence property allows for the generation of samples from f. The stationary distribution f can be specified to be almost any distribution, in

particular f can be chosen to be $p(\theta|y)$. The sequence $\mathbf{S}$ then converges to be a sample from $p(\theta|y)$. I use three MCMC algorithms to generate Markov chains. These are the Gibbs sampler (Geman and Geman 1984, Gelfand et al. 1990 and Gelfand and Smith 1990), the Metropolis-Hastings algorithm (Metropolis et al., 1953, Hastings, 1970), and a combination algorithm by Raghunathan et al. (2001).

2.2.1 Gibbs Sampling

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)$ be a P-vector with joint probability density $p(\boldsymbol{\theta}) = p(\theta_1, \ldots, \theta_P)$. The joint conditional density $p(\boldsymbol{\theta}|y)$ can be decomposed into its full conditional densities

$$\begin{aligned} &p(\theta_1|\theta_2, \ldots, \theta_P, y), \\ &p(\theta_2|\theta_1, \theta_3, \ldots, \theta_P, y), \\ &\quad\vdots \qquad (2.5) \\ &p(\theta_p|\theta_{-p}, y), \\ &\quad\vdots \qquad (2.6) \end{aligned}$$

where

$$\theta_{-p} = (\theta_1, \ldots, \theta_{p-1}, \theta_{p+1}, \ldots, \theta_P), \tag{2.7}$$

and where $\theta_p^{(\ell)}$ is the value sampled from density $p(\theta_p|\theta_{-p}^{(\ell)}, y)$ at the ℓ^{th} iteration and

$$\theta_{-p}^{(\ell)} = (\theta_1^{(\ell)}, \ldots, \theta_{p-1}^{(\ell)}, \theta_{p+1}^{(\ell-1)}, \ldots, \theta_P^{(\ell-1)}). \tag{2.8}$$

Algorithm 2.1 describes how Gibbs sampling works across iterations $\ell = 1, \ldots, L$, by sampling one parameter θ_p at a time from the parameter's full conditional density $p(\theta_p|\theta_{-p}^{(\ell-1)}, y)$ conditional on the current value of the other parameters. Sampling all $\theta_p, p = 1, \ldots, P$ in sequence constitutes P steps and one full iteration of the Gibbs sampling algorithm.

2.2.2 Metropolis-Hastings

If step p from equation (2.9) cannot be sampled from easily, a Metropolis-Hastings (M-H) algorithm can be employed in place of the Gibbs step. The Metropolis-Hastings algorithm

Algorithm 2.1 Gibbs Sampling Algorithm.

0. Choose a sensible starting value for $\theta_p^{(0)}, p = 1, \ldots, P$.

1. At iteration ℓ and step p, sample $\theta_p^{(\ell)}$ from its full conditional density conditional on the current value of the other parameters

$$p(\theta_p|\theta_{-p}^{(\ell-1)}, y), \tag{2.9}$$

where

$$\theta_{-p}^{(\ell-1)} = (\theta_1^{(\ell)}, \ldots, \theta_{p-1}^{(\ell)}, \theta_{p+1}^{(\ell-1)}, \ldots, \theta_P^{(\ell-1)}). \tag{2.10}$$

2. Repeat Step 1 (Gibbs step) for $p = 1, \ldots, P$.

3. Set $\boldsymbol{\theta}^{(\ell)} = (\theta_1^{(\ell)}, \ldots, \theta_P^{(\ell)})$.

4. Repeat Steps 1 and 3 for $t = 1, \ldots, T$.

is an alternative method of sampling from a target distribution $p(\theta_p|\theta_{-p}^{(\ell-1)}, y)$ by means of generating *candidate* values from a *proposal* distribution q, and then either accepting or rejecting the candidate values with *acceptance probability* ρ. Step p then is referred to as an M-H step, or Metropolis-within-Gibbs sampling. The general structure of this method is described in algorithm 2.2 and replaces the Gibbs step (step 1) of algorithm 2.1.

The random walk M-H algorithm is a popular choice for q. The candidate value is drawn as

$$\theta_p^* = \theta_p^{(\ell-1)} + \epsilon, \tag{2.11}$$

where ϵ is drawn from a symmetric proposal distribution q with mean zero and specified variance Σ. The candidate value θ^* then is in the neighborhood of the previous value in the Markov chain. When q is symmetric the acceptance probability ρ is

$$\rho = \min\left\{1, \frac{p(\theta_p^*|y)}{p(\theta_p^{(\ell-1)}|y)}\right\}. \tag{2.12}$$

Employing a good proposal variance $\boldsymbol{\Sigma}$ can improve algorithm efficiency tremendously.

Algorithm 2.2 Metropolis-Hastings Algorithm.

1. At iteration t generate a candidate value θ_p^* from a proposal distribution $q(\theta_p^*|\theta_p^{(\ell-1)})$.

2. Set

$$\theta_p^{(\ell)} = \begin{cases} \theta_p^* \text{ with probability } \rho \\ \theta_p^{(\ell-1)} \text{ with probability } 1-\rho, \end{cases}$$

where the *acceptance probability* ρ is

$$\rho = \min\left\{1, \frac{p(\theta_p^*|y)q(\theta_p^{(\ell-1)}|\theta_p^*)}{p(\theta_p^{(\ell-1)}|y)q(\theta_p^*|\theta_p^{(\ell-1)})}\right\}.$$

A good choice for $\boldsymbol{\Sigma}$ is $c\tilde{\boldsymbol{\Sigma}}$ where c is a variance scaling factor and $\tilde{\boldsymbol{\Sigma}}$ is an estimate of the posterior covariance matrix. This provides the algorithm with a measure of variance and covariance between the variables. The variance scaling factor c is used to adjust the magnitude of $\tilde{\boldsymbol{\Sigma}}$ to achieve a suitable acceptance rate near 25% (Gelman et al., 2004).

2.2.3 Adaptive Metropolis-Hastings Sampling

Allowing the proposal variance $\tilde{\boldsymbol{\Sigma}}$ to change during the simulation helps improve convergence of the MCMC chains (Müller, 1991). I follow the recommended strategy for posterior simulation outlined by Gelman et al. (2004, p. 307), in conjunction with re-estimation of the proposal variance similar to Müller (1991). The simulation is split into 2 phases. In the first phase the scaling factor c is allowed to change and the proposal variance $\tilde{\boldsymbol{\Sigma}}$ is re-estimated D times. Convergence of the MCMC chain is monitored during this phase. The second phase fixes c and $\tilde{\boldsymbol{\Sigma}}$ at their current values and continues for M_2 iterations.

Figure 2.1 provides a graphical representation of this adaptive sampling procedure. The parameters D, k, L_0, M_1, and M_2 are tuning parameters that are adjusted to the specific problem. The first phase is referred to as the *burn-in* phase. Samples from this phase are discarded. During this phase the first L_0 iterations are split into D blocks of equal length

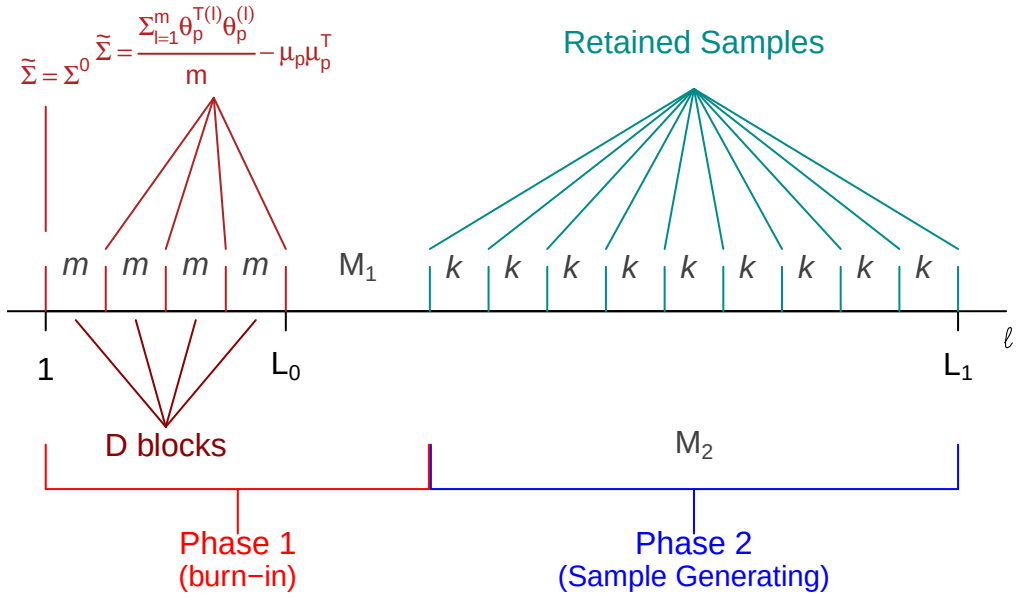


Figure 2.1: Diagram of the adaptive Metropolis-Hastings sampling algorithm.

m and indexed by $d = 1, \ldots, D$. At $l = 1, \tilde{\mathbf{\Sigma}}$ is initialized as $\tilde{\mathbf{\Sigma}}^0$, usually the identity matrix of appropriate dimension. At the termination of each block when iteration $l = dm$, $\tilde{\mathbf{\Sigma}}$ is re-estimated using the previous m samples of θ_p

$$\tilde{\mathbf{\Sigma}} = \frac{\sum_{\ell=1}^{m} \theta_p^{(\ell)} \theta_p^{'(\ell)}}{m} - \mu_p \mu_p^{'}, \tag{2.13}$$

where μ_p is the posterior mean of θ_p. The MCMC chains used in the following block $(d+1)$ are initialized as the values from the last iteration of the previous block (d). The value of c is initially set equal to 1, and is reset to 1 after each re-estimation of $\tilde{\mathbf{\Sigma}}$. How it is used in this adaptive phase is discussed next.

To achieve a suitable acceptance rate near 25% (Gelman et al., 2004) the variance scaling factor c is modified at each iteration in the following manner as suggested by Müller (1991). Let

$$\tilde{\rho} = \sum_{t-20}^{t-1} \rho_l \tag{2.14}$$

be the average acceptance probability over the 20 previous iterations. If $\tilde{\rho} > 0.5, c$ is increased by 20%, $c^{(\ell)} = 1.2c^{(\ell-1)}$, increasing the proposal variance, decreasing the probability of

accepting a new candidate value. If $\tilde{\rho} < 0.1, c$ is reduced by 30%, $c^{(l)} = 0.7c^{(\ell-1)}$ shrinking the proposal variance, increasing the chance the candidate value will be accepted.

When $t = L_0$, c and $\tilde{\boldsymbol{\Sigma}}$ are fixed at their current values and an additional M_1 iterations are run which serve as a final burn-in. The end of this burn-in marks the beginning of the second phase where the simulations are retained to build a sample from the target distribution. This phase consists of M_2 more iterations, assuming there is adequate mixing after M_1 iterations. To reduce the computational burden of post-simulation processing on extremely long chains and to reduce the correlation between sequential retained simulations only draws from the k^{th} iteration are retained in the final posterior sample resulting in a final sample size of $\tilde{n} = \frac{M_2}{k}$. This process is referred to as *thinning*. A recent discussion by Link and Eaton (2012) notes that the precision of the posterior estimates from the unthinned chain are more precise than those from the thinned chain. Christensen et al. (2010, p. 146) directly states that unless the autocorrelation by lag 30 is still very high, thinning isn't worthwhile. In this dissertation I use a thin large enough such that the retained sample is mostly uncorrelated and that will result in nice round final sample size of 1,000 or 5,000 for example.

2.3 Monitoring and Assessing Convergence

I take a heuristic approach to determining if the chain has converged by monitoring several diagnostic measures suggested in Albert (2009) and provided in the `coda` package in `R`.

- Trace plots: When the trajectory of $\theta_p^{(\ell)}$ levels off such that the random oscillations remain within bounds the chain is considered to have converged. This plot is also used to ensure the chain is exploring the parameter space sufficiently.

- Acceptance rate: This is monitored throughout the burn-in period and at the end of the second phase.

- Autocorrelation plots are inspected to determine the serial correlations of the samples, which helps determine a thinning fraction $1/k$.

- Density plots are inspected to assess the smoothness of the density of the retained sample. This helps to decide the length of the final chain.

- Gelman-Rubin-Brooks plots (Brooks and Gelman, 1998) are created when multiple chains are run to monitor the shrink factor as the number of iterations increases. A median shrink factor close to 1 is a positive indication of convergence.

2.4 Multiple Chains and Parallel Computing

Generating s multiple MCMC chains with diffuse starting values to estimate θ_p is advantageous for several reasons. They can help ensure the entire parameter space is fully explored, and that all chains converge to the same region. They can also cut down the number of simulations needed per chain, one chain of $T = 10,000$ can be broken down into s chains of length $T = 10,000/s$ each. Unless otherwise specified I generate somewhat diffuse starting values for the multiple MCMC chains by choosing sensible starting values for the first chain, then adding random deviates to each value and using those as the starting values for the subsequent chains. For example let the starting values for the first MCMC chain for $\boldsymbol{\theta}$ be (a_1, a_2) and let u_1 and u_2 be draws from a $U(-1, 1)$ distribution. The starting values for the second MCMC chain then would be $(a_1 + u_1, a_2 + u_2)$.

Most MCMC algorithms can also benefit from parallel computing. Parallel computing, is when the tasks assigned by the algorithm are split across several CPU's (or computers). Standard programming utilizes a single CPU and runs one chain at a time sequentially. Parallel programming can assign the computation of s different chains to different CPU's. For example if you specify to use $s = 4$ chains and 4 CPU's, each chain can be run on its own CPU. If you ask for $s = 8$ chains on 4 CPU's, each CPU will run 2 chains sequentially. This does not guarantee that the computation time is divided by s, but it usually provides a significant reduction.

The adaptive random walk M-H sampling algorithm lends itself reasonably well to parallel processing thanks to the `snowfall` package (Knaus, 2010) in `R` (R Core Team, 2012).

`snowfall` is a user friendly wrapper for the `SNOW` (Simple Network of Workstations) package (Tierney et al., 2011) for `R` that controls the workflow from the single `R` instance to multiple CPUS's. I walk through a very generic example of how a function that performs MCMC sampling can be written as a parallel process. First I provide code for a single chain on a single CPU, then extend this to multiple chains on multiple CPU's running in parallel.

Code example The goal for this example is to sample from the posterior distribution $p(\theta|y)$ by creating an MCMC chain using a function `F()`. This function takes data, priors, and simulation parameters as inputs, and generates a single MCMC chain that estimates the desired posterior distribution. For simplicity and to stay focused on the parallel processing aspects of the example I do not go into the sampling algorithm details and just use `F()` to sample from the desired posterior distribution.

Let parameter θ have prior density with mean `mu` and precision `P`. The simulation is run for `t`=1100 iterations, with a burn-in of $b = 100$ and retaining every $k = 5^{\text{th}}$ iteration. Additional arguments `y` provide data, and `start` gives starting values for the MCMC chain. The function call is

```
single.chain <- F(t = 1100, b = 100, k = 5,
    y = y, start = t0, prior = list(mean = mu, prec=P)).
```

The returned object `single.chain` is a vector of length $\tilde{n}$ =(1100-100)/5= 200. To utilize multiple chains and multiple processors, first each CPU has to be initialized and linked to `R` (`sfInit`). Then all data, code and packages are loaded onto each CPU (`SfExport`). Once the random number generator on each CPU is initialized separately (`sfClusterSetupRNG`), the function `F()` can be called s times using `sfLapply`, the `snowfall` version of `lapply`.

```
# Define number of chains to use
  s  <- 4
# Create diffuse starting values
  t1 <- matrix(rep(t0, s), ncol=s) + runif(n=length(t0)*s, min=-1, max=1)
```

```
# Gather the CPU's
  sfInit(parallel=TRUE, cpus=4)
# Export data to all CPU's
 sfExport("y", "t1", "prior")
# Set random seed generator on each cluster
  sfClusterSetupRNG()
# Apply the function F() to each of 1:s chains on each of 1 to 4 CPU's
parallel.sim <- sfLapply(1:s, function(x){
  F(t = 1100, b = 100, k = 5, start = t1[,x],
  prior=list(mean=mu, prec=P)})
# Release the CPU's
  sfStop()
```

After each chain completes, the s resulting objects are returned as a list into the object `parallel.sim`. Each of the 4 items in the list are single MCMC chains of length 200, creating a total posterior sample size of $\tilde{n} * s = 1,000$. At this point each CPU that was gathered to be used in this simulation needs to be released, or unlinked, from `R` using `sfStop`.

Figure 2.2 displays a selection of computation times between parallel and sequential simulations. The advantage of parallel processing is clear as early as 1,000 iterations. Running 1,000 iterations on each of 4 chains takes 18.4 minutes when using a single CPU and 15.6 minutes when using 4 CPU's. This difference of about 3 minutes increases to a 10 minute gain when running 2,000 simulations. Even this relatively small number of iterations takes 43.5 minutes for a single CPU and 33.4 minutes for 4 CPUs.

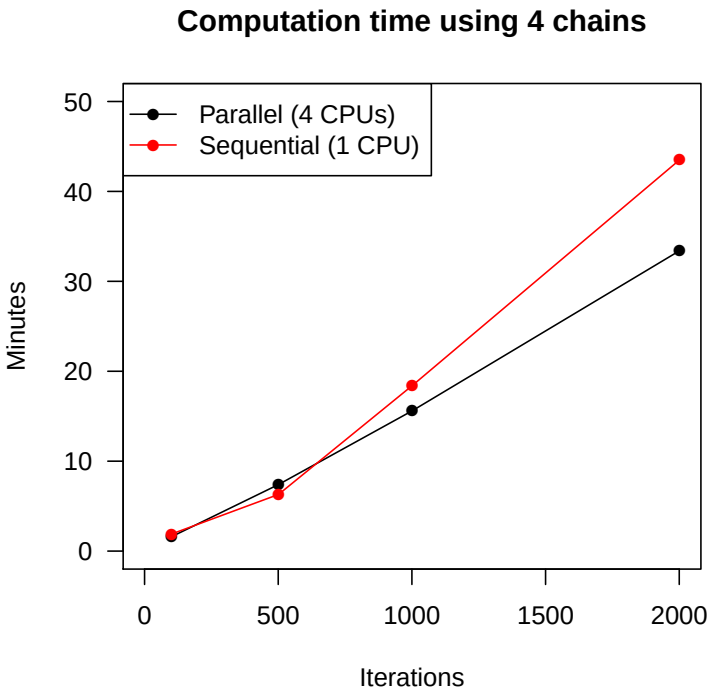


Figure 2.2: Comparison of computation time between sequential and parallel processing when using 4 chains on Intel Core i7 920 2.67GHz, 12G RAM.

CHAPTER 3

Review of Editing and Imputation Procedures

This chapter discusses the concepts of editing and imputation, including historical treatment and some current procedures commonly used in data editing. I describe five primary frameworks used to impute missing data and some commonly used terms within the missing data literature. I present a recent literature review on imputation and editing and then describe some techniques for and advantages of combining the editing and imputation procedures. Some notation used throughout the dissertation is introduced.

3.1 Imputation

Imputation is now a standard technique for handling missing data (Little and Rubin, 2002). Rubin (1976) coined terms that classify the relationship between the missingness, which is the process resulting in missing values, and the missing and observed values themselves. Let y be a vector of data from a single subject that contains both missing and observed values. Without loss of generality, the data vector y can be split into y^{mis} and y^{obs}, the sub-vectors of y that are missing and observed respectively. Let X be any other fully observed variables of interest. The possible mechanisms are

- Missing Completely at Random (MCAR): The probability that components of y are missing is unrelated to any other observed data y^{obs}, X or the unobserved y^{mis} value. This assumption can be partially tested (Little, 1988).
- Missing at Random (MAR): This is also called *ignorable*. The probability that y is missing is dependent only on the observed values y^{obs} and X, but not on the values of the missing data itself.

- Not Missing at Random (NMAR): The probability that y is missing is dependent on y^{mis} as well as possibly y^{obs} and X.

Most standard imputation methods depend on the MAR assumption. Reiter and Raghunathan (2007) suggest that the applicability of this assumption is related to the information content of other related variables that can be used in an imputation process. Rancourt (2001) describes the following five main frameworks used to impute data and which are useful for understanding the mechanisms of how a value is imputed. Rancourt's taxonomy is:

1. *Experience-based*: Experts are performing imputations based on their knowledge.
2. *Distribution-based*: Distributions are estimated and imputations are obtained from them.
3. *Model-based*: A model is constructed, validated and used to produce values to impute.
4. *Frequency-based*: Under a response mechanism, values are imputed without using a model.
5. *Empirical-based (donor-based)*: A donor is found and its values are used for imputation.

Regardless of how the imputation is performed, single imputation does not account for the error incurred by the imputation process itself. Treating the imputed values as if they were known can lead to inappropriately smaller variances and therefore an erroneous increased chance of significant findings.

3.1.1 Multiple Imputation

One common way of adjusting the analysis to include the error associated with filling in missing values is multiple imputation (MI) (Little and Rubin, 2002). The idea, as graphically presented in Figure 3.1, is to perform an imputation procedure containing a random component multiple times, creating multiple complete data sets. The desired estimate from the analysis, such as a mean or regression coefficient, is then calculated on each data set

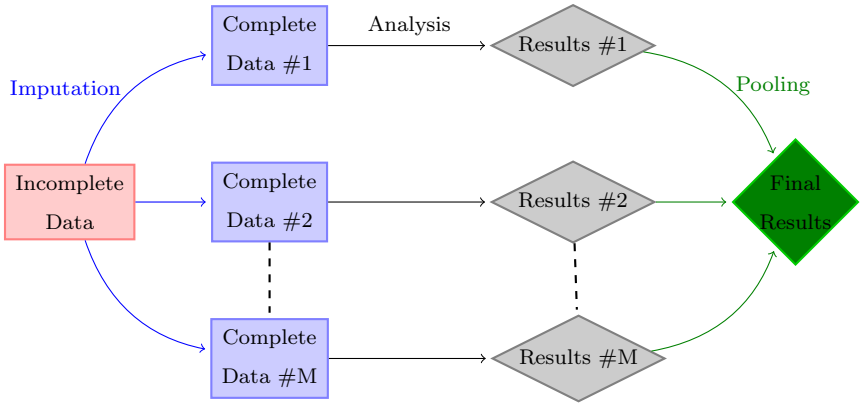


Figure 3.1: Graphical representation of the multiple imputation process.

separately. The estimates are pooled using simple combining rules, also known as *Rubin's Rules* (Rubin, 1987) and are defined as follows.

Let δ be the parameter whose estimate we desire to obtain from an analysis. Given M imputed data sets, M estimates of δ : $(\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_M)$ are generated and used to calculate the following quantities.

- The overall estimate is the average of the individual point estimates

$$\hat{Q} = \frac{1}{M} \sum_{m=1}^{M} \hat{\delta}_m. \tag{3.1}$$

- The within-imputation variance is the average of the individual variances

$$U = \frac{1}{M} \sum_{m=1}^{M} Var(\delta_m). \tag{3.2}$$

- The between-imputation variance is the variance of the estimates

$$B = Var(\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_M). \tag{3.3}$$

- The total variance is

$$T = U + (1 + \frac{1}{M})B, \tag{3.4}$$

- and 95% intervals are calculated as

$$\hat{Q} \pm 1.96 * \sqrt{T}. \tag{3.5}$$

The resulting variance of the combined estimate then accounts for both the within and between data set variances. Reiter and Raghunathan (2007) give a review for when these rules are valid and how they should be corrected under certain circumstances. The methodology introduced in this dissertation is a situation where Rubin's Rules are valid.

Raghunathan et al. (2001) argue that the best or most appropriate framework for performing multiple imputation is under a fully Bayesian model where a model for the missing data conditional on the observed data, prior distributions for all parameters, and a model for the missing data mechanism are all explicitly defined. Rubin (1987) introduced the general framework to generate multiple imputations for the missing values y^{mis} given the data generating model parameter θ.

1. Calculate the conditional posterior density $P(\theta|y^{obs})$ of the model parameters θ conditional on the observed data.

2. Sample a value θ^* from $P(\theta|y^{obs})$.

3. For each observation in y^{mis}, y^* is drawn from $P(y^{mis}|y^{obs}, \theta^*)$, the predictive distribution for the missing values given the sampled parameter values.

Steps 2-3 are repeated multiple times to create multiple imputations. This method has been extended to impute missing values under various multivariate distributions (Rubin and Schafer, 1990, Schafer, 1997).

One method to create multiple multivariate imputations for a complex model under a Bayesian framework is called Sequential Regression Multivariate Imputation (SRMI) (Raghunathan et al., 2001). This method presents a strategy to create multiple imputations by drawing from the posterior predictive distribution of the missing data using a multivariate regression model for the variables being imputed using all other variables as predictors. Imputations are done sequentially on a variable by variable basis, and is cyclical in that newly imputed values overwrite the previous imputed values. The multiple imputation and editing procedures I introduce in this dissertation build off this technique, and the next section introduces some notation and describes how SRMI works in detail.

3.2 Notation

The notation introduced here assumes no repeated measures in the data set. Chapter 4 expands this notation to accommodate longitudinal data.

Let Y_p, $p = 1, \ldots, P$ represent P variables that are subject to missing data. Then $Y = (Y_1, \ldots, Y_P)^{'}$ is the full collection of Y_P variables, and $Y_{-p} = (Y_1, \ldots, Y_{p-1}, Y_{p+1}, \ldots, Y_P)^{'}$ is the collection of all variables subject to missing data *except* Y_p. Let y_{ip} be the i^{th} subject's p^{th} outcome, $i = 1, \ldots, n$, $\mathbf{y}_p = (y_{1p}, \ldots, y_{np})^{'}$. Then $\mathbf{y}_{n\times P} = (\mathbf{y}_1, \ldots, \mathbf{y}_P)$ is the data matrix of all Y and $\mathbf{y}_{-p} = (\mathbf{y}_1, \ldots, \mathbf{y}_{p-1}, \mathbf{y}_{p+1}, \ldots, \mathbf{y}_P)^{'}$ is the $n \times (P-1)$ data matrix for all variables *except* $\mathbf{y}_p$. Let $X = (X_1, \ldots, X_Q)^{'}$ be the collection of all Q fully observed variables with x_{iq} the data from observation i on variable Y_q. Then $X_q = (x_{1q}, \ldots, x_{nq})^{'}$ the vector of data for variable X_q, and $\mathbf{x}_{n\times Q} = (\mathbf{x}_1, \ldots, \mathbf{x}_Q)^{'}$ is the matrix of fully observed data with $\mathbf{x}_i^{'} = (x_{i1}, \ldots, x_{iQ})$ in the ith row.

Let each vector Y_p have an associated vector M_p of missing data indicators. For example if Y_2 has missing values then $M_2 = (M_{12}, \ldots, M_{n2})^{'}$ where $M_{i2} = 1$ if y_{i2} is missing and 0 otherwise. The count of missing observations for Y_p is $n_p^{miss} = \sum_i M_{ip}$ for all $p = 1, \ldots, P$. It follows that $n_p^{obs} = n - n_p^{miss}$ is the count of data $y_{ip}, i = 1, \ldots, n$, that is observed.

3.2.1 Sequential Regression Multivariate Imputation

The Sequential Regression Multivariate Imputation (SRMI) (Raghunathan et al., 2001) procedure defines a regression model g_p for each variable Y_p that is specific to the variable type of Y_p. For example a logistic regression for binary Y_p, linear regression for continuous Y_p or Poisson log-linear model for a count variable Y_p. In general

$$Y_p \sim g_p(Y_{-p}, X, \boldsymbol{\theta}_p), \tag{3.6}$$

where $\boldsymbol{\theta}_p$ is a vector of parameters specific to the density g_p, which could include regression coefficients and variance or dispersion parameters. The SRMI procedure uses only the observed Y_p cases to fit the models, in other words all rows where $M_{ip} = 0$. A sample $\boldsymbol{\theta}_p^*$ is drawn from the posterior density of $\boldsymbol{\theta}_p$ conditional on the data used in fitting the regression

model, and subsequently used to draw imputed values from the distribution of the missing y_{ip} where $M_{ip} = 1$ given the observed data and $\boldsymbol{\theta}_p^*$.

Specifically at iteration t,

1. The observed values for y_{ip} are regressed on the most recently updated version of $\mathbf{y}_{-p}^{(\ell)}$ which consists of the complete (observed and imputed from the current iteration) data $\mathbf{y}_1^{(\ell)}, \ldots, \mathbf{y}_{p-1}^{(\ell)}$, complete (observed and imputed from the previous iteration) data for $\mathbf{y}_{p+1}^{(\ell-1)}, \ldots, \mathbf{y}_P^{(\ell-1)}$, and the fully observed data X.

2. A sample value $\boldsymbol{\theta}_p^{(\ell)}$ is then drawn from the posterior density of $\boldsymbol{\theta}_p$ conditional on observed y_{ip}, and $y_{-p}^{(\ell)}$,

$$p(\boldsymbol{\theta}_p|\mathbf{y}_p, \mathbf{y}_1^{(\ell)}, \ldots, \mathbf{y}_{p-1}^{(\ell)}, \mathbf{y}_{p+1}^{(\ell-1)}, \ldots, \mathbf{y}_P^{(\ell-1)}, X). \tag{3.7}$$

3. For all n_p^{mis} observations with $M_{ip} = 1$, imputed values $y_{ip}^{I(\ell)}$ are drawn from

$$g_p(y_{ip}^I|\boldsymbol{\theta}_p^{(\ell)}, y_{i1}^{(\ell)}, \ldots, y_{ip-1}^{(\ell)}, y_{ip+1}^{(\ell-1)}, \ldots, y_{iP}^{(\ell-1)}, \mathbf{x}_i), \tag{3.8}$$

   These drawn values are used to impute the missing y_{ip} to create a fully complete vector $\mathbf{y}_p^{(\ell)}$.

This algorithm cycles through all $Y_1, \ldots, Y_p$ variables in each iteration. Multiple imputations can be created by retaining values from every k^{th} iteration, or by using a single draw from M parallel chains with diffuse starting values. These retained values can be merged back into the original data set to create multiple complete data sets on which analysis can be performed.

Raghunathan et al. (2001) point out that since the P conditional distributions from (3.8) do not necessarily derive from a full joint distribution there is no theoretical guarantee of convergence to a stationary distribution. However based on their test data sets and subsequent research they do not feel this is an actual problem. SRMI has been available to researchers since 1997 (last updated in 2011) in the computer software program IVEWARE (stand alone and available for use in SAS) with no indication that this lack of confirmed

theoretical convergence is an issue. A similar technique by van Buuren et al. (1999) called Multiple Imputation using Chained Equations (MICE) also performs variable by variable imputation using distinct conditional distributions. This procedure has been incorporated into the main STATA distribution (Royston, 2004, Royston and White, 2011) and R (van Buuren and Groothuis-Oudshoorn, 2011). Cyclical procedures such as SRMI and MICE are useful techniques for performing multivariate MI on mixed data types.

Extensions to multiple imputation. The original MI process has been examined, modified and enhanced as the range of statistical problems have become increasingly complex (Reiter and Raghunathan, 2007). Here I discuss enhancements that deal with missing data in longitudinal studies, and some studies where imputation and editing are combined.

Many modifications to the MI framework are done under the auspice of an ignorable missing data mechanism. It is not always appropriate to assume that missingness in a longitudinal study is ignorable, for example missingness due to drop out. Drop out causes all data from an individual after a certain point in time to be missing. This differs from intermittent missingness where a participant is still in the study but does not contribute data at a given time point. Without further study details, simply examining the missing data pattern cannot differentiate between intermittent missingness that looks like dropout, and direct dropout. Yang and Shoptaw (2005) describe an imputation framework called Multiple Partial Imputation for longitudinal studies with intermittent missingness and dropout. This is a two step process to impute the intermittent missingness first, and then deal with the dropout separately.

Other techniques to handle missing data in longitudinal studies have been explored by Yang et al. (2008), Demirtas and Hedeker (2007, 2008), and Daniels and Hogan (2007). Tang et al. (2005) provides a comparison of imputation methods for longitudinal studies. Several review papers including Reiter and Raghunathan (2007), Raghunathan (2004), Horton and Kleinman (2007) have summarized the missing data concepts, techniques and software to date. Yucel (2011) provides a recent review of the state of multiple imputation software programs that exist as stand alone software or have been incorporated into a variety of

statistical software packages. Practitioners interested in multiple imputation have many options to choose from; the same can not be said about editing procedures.

3.3 Editing

Data editing is done to correct outlying values, values considered to be in error, and inconsistent combinations of responses. Edits to the data are also called recodes, the terms "edit" and "recode" are used interchangeably. These recodes typically have the form of a series of logical constraints and deterministic if-then type rules.

Historically edits were performed by people manually going over each record and checking for internal consistency in accordance with edit rules (Herzog et al., 2007). These rules could involve one or more of the following outlier detection methods. A univariate distributional analysis where values outside a specific range are considered implausible (monthly income of $1 million or more) or impossible (observing a value of 2 when the response options are only 0 or 1), or a bivariate distributional analysis where the responses of two continuous variables are plotted against each other and outlying points are identified. The outlier detection method most appropriate for multivariate categorical data is a multivariate pattern that the science or experts dictate is impossible or highly unlikely and therefore could be erroneous (15 year old reporting having a 30 year old child).

Computers have aided this editing process by increasing the speed and consistency of these recodes, removing the burden on the individual performing the edit checks, but the software to perform these recodes can be difficult to write (Herzog et al., 2007). Large scale surveys can have hundreds of these recode rules that require an iterative editing process to ensure all rules are satisfied, and still may require final personal review and manual editing.

Fellegi and Holt (1976) developed a system that would ensure all records satisfy all recode rules in one pass of the data through the editing process. Their work comes out of Operations Research and gained greater notice by business statisticians and econometricians rather than Biostatisticians in Public Health. de Waal and Coutinho (2005) provide a review of the Fellegi-Holt method and compare it to three other algorithms to locate and edit errors in

business surveys.

3.4 Combining Edit & Imputation

Several programs exist that combine various single edit and single imputation (E&I) procedures. The annual Conference of European Statisticians (2011) held a work session on Statistical Data Editing. Statisticians working on government level censuses and surveys met to discuss new developments in their countries E&I systems. However these programs are very large scale and designed for use on the governmental level and are not easily available to the individual researcher.

Examples of these programs include the Structural Programs for Economic Editing and Referrals (SPEER) (Winkler and Draper, 1996) and DISCRETE (Winkler and Petkunas, 1997) editing systems used by the US Census Bureau, the Generalized Edit and Imputation System (GEIS/Banaff) used by Statistics Canada (Kovar et al., 1991), and the program Data Imputation Editing System - Italian Software (DIESIS) (Bruni et al., 2002) used by the Italian government. These programs are designed to handle very large scale surveys, require a decent amount of manpower to implement, and are rather general and must be configured for each different survey (Herzog et al., 2007, Ghosh-Dastidar and Schafer, 2003).

The EUREDIT (2003) project compared many editing and imputation procedures on the quality of the imputed values, but did not specifically examine these programs for their ability to impute values satisfying prespecified editing rules, nor their ability to incorporate the additional uncertainty introduced by the editing and imputation processes themselves. They specifically state that "evaluation of MI versions of the EUREDIT imputation methods would require a large-scale simulation exercise..." which was outside the scope of the EUREDIT project.

Limited publications exist which develop methods of combining editing and imputation procedures that account for the additional variance introduced by the E&I procedure itself. Ghosh-Dastidar and Schafer (2003) combined multiple editing with multiple imputation using a Bayesian approach that can be used for continuous data with intermittent errors.

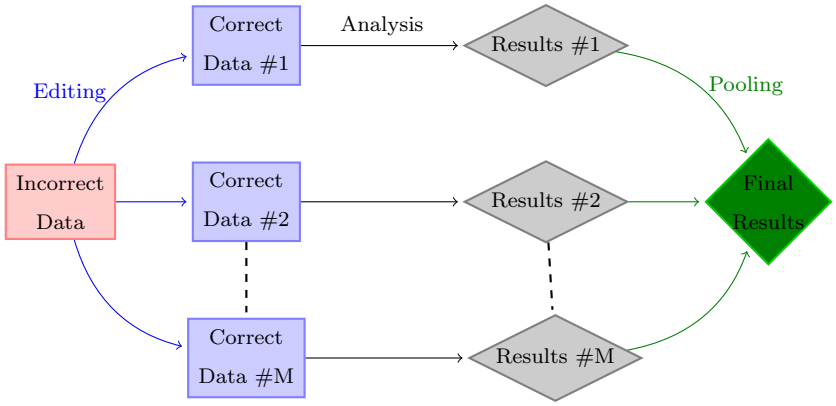


Figure 3.2: Graphical representation of the multiple editing process.

Winkler (2003) introduced a method to edit inconsistent categorical data by first blanking out (making the values missing) all inconsistent responses, then imputing all missing values using a generalized hot-deck imputation method that imputes values which satisfy pre-specified editing constraints. This method is based on the Fellegi and Holt theory of editing and so requires some knowledge of operations research to implement. Cole et al. (2006) provide a multiple imputation for measurement error (MIME) correction, but commentator White (2006) pointed out this method is only appropriate when the true response values for a subset of observations is available.

The common theme with multiple editing processes is that the edited values are no longer deterministically assigned, but are a result of a regression prediction or sampled from a posterior distribution. Figure 3.2 shows that the process of multiple editing is the same concept as multiple imputation. No work has been identified that uses the concept of multiple editing in a longitudinal study.

3.4.1 Expansion of Notation to Accommodate Inconsistent Data

To accomplish my goal of introducing a procedure that combines stochastic editing processes and multiple imputation, some notation needs to be expanded to include editing indicators. First I expand the definition of Y_p to include variables that are subject to missing and/or erroneous data.

Let $E_k = (E_{1k}, \ldots, E_{nk})'$ be a vector of editing indicators for the k^{th} known edit where $E_{ik} = 1$ if record i is to be edited and $E_{ik} = 0$ otherwise. While there is a one to one connection between Y_p and M_p, a single editing indicator E_k could represent a multivariate edit and thus correspond to an inconsistent pair of, say, Y_k and Y_{k+1}.

To better explain how the missing and editing indicators work with the data, consider three binary variables Y_4, Y_5, Y_6 with valid values of $(0, 1)$, where Y_4 and Y_5 contain some missing values (**NA**). In addition, variable Y_4 has some invalid cases with values of $y_4 = 2$, and the combination $(Y_5, Y_6) = (1, 1)$ is inconsistent and is to be edited. Two edit indicators are defined as $E_1 = \mathbf{1}\{y_4 = 2\}$ and $E_2 = \mathbf{1}\{y_5 = 1 \cap y_6 = 1\}$. An example first four rows of a matrix containing (Y_4, Y_5, Y_6) and its associated imputation and editing vectors of indicators might look like

$$
\begin{array}{c}
i \\ 1 \\ 2 \\ 3 \\ 4
\end{array}
\begin{array}{c}
\begin{array}{ccccccc} y_4 & y_5 & y_6 & M_4 & M_5 & E_1 & E_2 \end{array} \\
\left(\begin{array}{ccccccc}
0 & \mathbf{NA} & 0 & 0 & \mathbf{1} & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 \\
\mathbf{NA} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & \mathbf{1} \\
\mathbf{2} & 0 & 0 & 0 & 0 & \mathbf{1} & 0
\end{array}\right)
\end{array}.
$$

In the first row, y_{i5} is missing ($y_{15} = \mathbf{NA}$) and so $M_{51} = 1$. Similarly y_{34} is missing, so $M_{34} = 1$. Also $E_{41} = 1$ because $y_{44} = 2$, and $E_{32} = 1$ because the pair $(y_{35}, y_{36}) = (1, 1)$. Further let $n_k^{err} = \sum_i E_k$ be the number of observations to be edited under edit rule k.

3.5 Data Editing and Imputation in the Project Connect data

Project Connect used an experience-based editing and imputation framework, executed as a series of if-then logical statements to perform edits and imputations on the survey data set. Inconsistent multivariable responses were edited each year as the data was collected, inconsistent repeated measures were edited after the final year of data collection. Editing a single year without regards to data collected on the same person in other years may result in inconsistencies across years. I discuss some of the general recode rules currently implemented in the PC data set and go into detail regarding some specific edits from the Sexual Activity

section (Section I).

3.5.1 Current Editing and Imputation Rules

Here are a few examples of the currently implemented editing and imputation rules. The full text of questions is given in Appendix A.

1. Demographics: Gender (`A1`), Birth date (`A2,A3`), Ethnicity (`A7`). Rules include single yearly imputations and longitudinal edits.

   (a) Yearly single imputations: If gender, birth date or ethnicity are missing during a single year, impute them with the school roster data.

   (b) Longitudinal edits: If the value of gender changes across years, use the value that is most often reported (majority rules) as the true value. Roster data is used as a tie breaker.

2. School Based Health Center (SBHC). Question `E1` (`KNOW SBHC`) asks if the student knows of an SBHC on their high school campus. Question `E2` (`EVER BEEN`) asks if the student has been to the SBHC. Question `E3` (`WHY`) asks what the student has visited the SBHC for, and has 12 parts (a-l) that list activities such as "Immunizations" or "Birth Control". Question `E4` (`WHY NOT`) has 8 parts, asking reasons why the student has **not** visited the SBHC, and contains several options such as "I didn't feel comfortable", and "I thought I'd have to pay".

   (a) If the school doesn't have an SBHC then change `E1-E3` to missing.

   (b) If there are conflicts (some Yes, some No) among `KNOW SBHC, EVER BEEN` and `WHY` variables then change all of them to missing.

   (c) If `KNOW SBHC & EVER BEEN` are missing, and `WHY` are all No then change `WHY` to missing.

   (d) If `KNOW SBHC` is not missing and `EVER BEEN` is missing and `WHY` is all No responses then change `EVER BEEN` to No.

(e) If KNOW SBHC is Yes and EVER BEEN is No and WHY has some Yes responses then change EVER BEEN to Yes.

(f) If KNOW SBHC and EVER BEEN are missing and WHY has some Yes then change KNOW SBHC and EVER BEEN to Yes.

3. Condom Availability Program (CAP): Question G1 (KNOW) asks about student awareness of the CAP. Question G2 (WHO) asks if specific people give out condoms on campus. These two questions have a "Don't Know" response option. Question G3 (UTIL) asks about condom acquisition from the CAP. Question G4 (AMOUNT) asks how many times the student has gotten condoms from the CAP in the past month.

(a) CAP Recode Step 1: Impute Missing

i. Change Don't Know to No for KNOW and all of WHO.

ii. If KNOW is No or Don't know and skipped the entire rest of the section, then change all WHO, and UTIL to No.

iii. If KNOW is missing but the student answered the rest of the section, and if they said Yes to at least one of WHO change KNOW to Yes.

(b) CAP Recode Step 2: Edit Inconsistencies

i. If they said Yes to 2 or more of WHO then change KNOW to Yes.

ii. If they said No to KNOW and Yes to UTIL then change UTIL to No.

3.5.2 Sexual Activity Data Editing and Imputation

Section I contains the questions that measure sexual risk behaviors. Missing data and inconsistencies in responses within this section need to be dealt with in a consistent and logical manner. A primary goal of the intervention was to postpone the time to first sexual intercourse. This was measured by examining the change across years in the proportion of students who reported ever having sexual intercourse, and by measuring the change in reported age of sexual onset. Any recode that alters variables involved in these measurements

has the potential to change the perceived effect of the intervention.

Currently the error patterns in this section are resolved by a series of order-specific deterministic recode rules. A total of 26 edit rules were applied to this section and are listed in Appendix Table A.5. The first 14 recode rules address inconsistencies and missing data in the first five questions, `I1-I5`. Briefly, `I1` asks about lifetime sexual experience, `I2-I3` ask the month and year of when the student first had sexual intercourse. Question `I4` asks the age the student was at first sex, and `I5` asks about the number of lifetime partners.

One way to visualize the complexity of these rules is depicted in Figure 3.3. The green hexagons represent the ***if*** conditions. The arrow from the green hexagon to a white box represents the ***and*** condition, and the arrow from the white box leads to a scroll that contains the editing decision (***then***). Red scrolls are recodes that change `I1` directly, blue scrolls are edits that change other variables. The missing data codes used are M =Missing, S ==I have never had sex, K ==Don't Know.

Let's walk through recode #1:*If `I1` is "M" and `I2` through `I5` are all "S" then change `I1` to No.* Recode #1 can be traced as follows: The green hexagon at the top of the page that says I1=M represents *If `I1` is "M"*. Following the arrow to the left to the white box that contains (1) I2-I5 ARE ALL S represents the *and `I2` through `I5` are all "S"* portion of the recode rule. The (1) indicates that it is Recode Rule #1. Following the arrow down to the red scroll that says CHANGE I1 TO NO instructs you to *then change `I1` to No* for observations where both of these conditions are satisfied.

Ten percent of observations during the first three years and 12% during the fourth year were affected by at least one edit in the sexual activity section. The main contributor is edit rule #25 which edits inconsistencies between having sex in the past 3 months and consistency in condom use during sex in the past three months. Edit rules such as these generate additional missing data that could impact analyses.

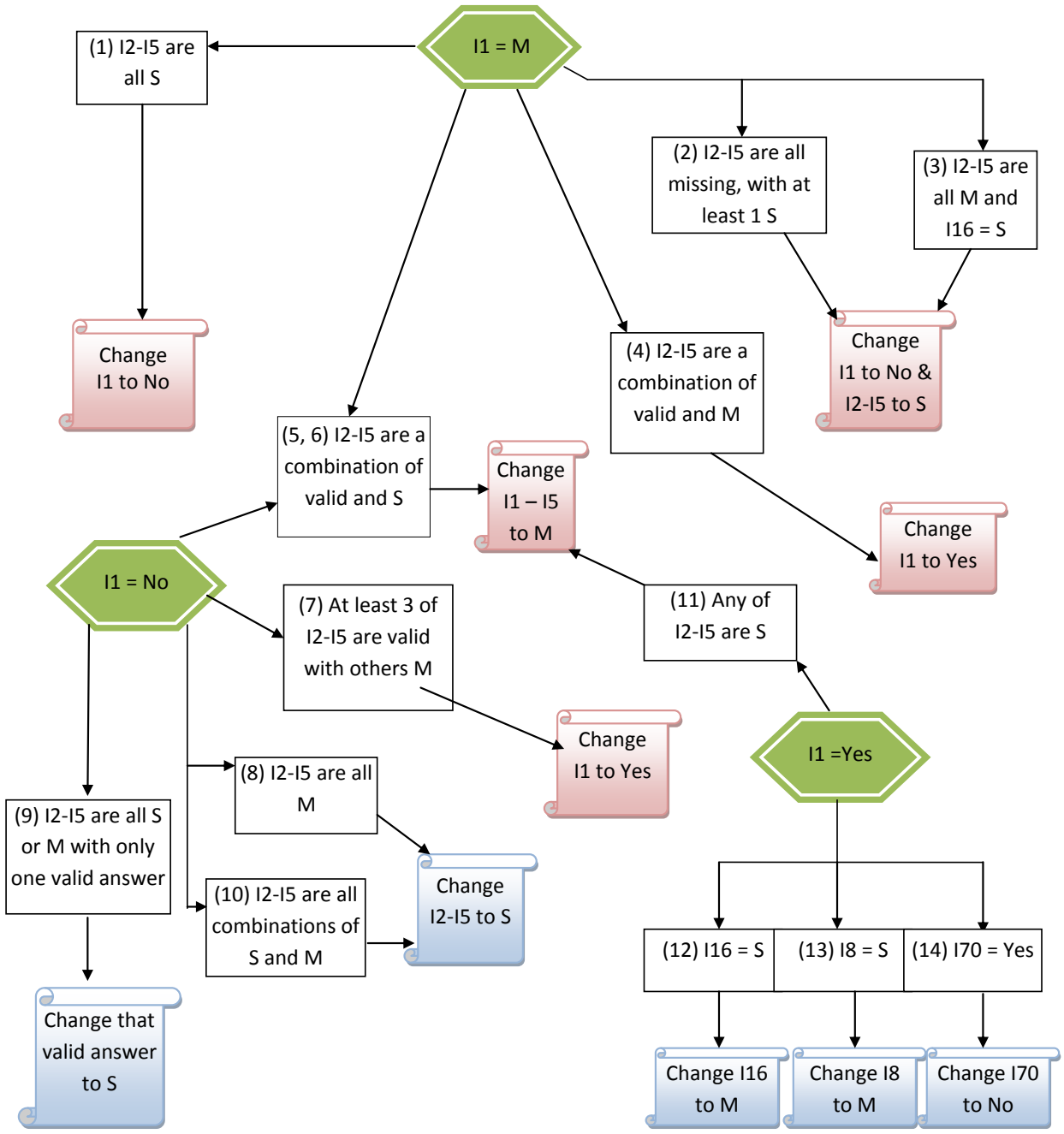


Figure 3.3: A graphical representation of the first 14 recode rules found in Appendix Table A.5. These edits revolve around the primary question of ever having sexual intercourse (`I1`). Special missing data codes: M=Missing, S=I have never had sex.

3.6 Conclusion

The major assumption when using a set of deterministic recode rules to change data is that the edits are considered to be 100% correct for every affected observation. These rules are usually generated on the basis of what the researcher assumes the participant "should have" or "meant to" report using their scientific knowledge of the situation and the observed reported values. Data gained by singly imputing missing data or data changed by using a deterministic editing process are treated in follow up analysis as if the data are without error.

I distinguish between survey reported values y_i and the underlying truth Z_i. Deterministic editing assigns a particular value to Z_i as a function of observed data y_i with probability 1. But often these rules are not thought to be perfect and we do not expect that the assumed Z_i is always correct due to mis-reporting, random errors or active lying in a subset of the population.

I develop Bayesian multiple editing methods to correct certain types of errors that occur in the Project Connect data set. Using the Bayesian paradigm allows for good scientific prior knowledge about the problem to be included in these models and encoded in the editing rules. Edited values are drawn from the posterior distribution of the correct data given the clearly incorrect responses. These methods correctly propagate error from the editing and imputation procedures into the final analysis and conclusions. These procedures result in multiply imputed and multiply edited data sets that can then be used for analyses.

CHAPTER 4

Modeling Time Fixed Inconsistent Repeated Measures

This chapter introduces a model and subsequent imputation and editing procedures for the case of inconsistent repeated measurements (IRM) of a time-fixed binary variable. There is a model for the subject's underlying true value, z_i, and a model for the reporting process, y_{ij}, $i = 1, \ldots, n, j = 1, \ldots, m_j$. The unknown latent variable z_i remains constant across time and has two possible values, either 0 or 1.

I use a Bayesian hierarchical latent variable model to estimate the underlying true value z_i given reported values y_{ij}. This model incorporates prior knowledge about the association of predictors with the latent characteristic. The probability that z_i is 1 is modeled with a logistic regression on a set of q predictors, $\mathbf{w}_i$. The reporting model for y_{ij} given z_i is a mixture model where the regression parameters differ based on the value of z_i. The probability that y_{ij} is 1 is modeled with a logistic regression on a second set of h predictors, $\mathbf{x}_{ij}$, which are associated with how a student reports the characteristic.

The models are fit with the Markov chain Monte Carlo (MCMC) simulation techniques described in Chapter 2.2. Draws from the resulting posterior sample are used in a multiple imputation and editing procedure to impute the missing and correct the inconsistent values. This chapter details the IRM model, posterior density calculations and sampling algorithm. I apply the IRM model twice, to model their true gender and to model participant's true birthplace.

4.1 Model Specification

The underlying true value of the characteristic of interest, z_i, is modeled as a Bernoulli random variable with probability λ_i. The logit of λ_i is assumed to be a linear combination of q subject level predictors, $\mathbf{w_i} = (w_{i1}, \ldots, w_{iq})'$ and corresponding q-vector of regression coefficients, $\boldsymbol{\gamma}$; $\text{logit}(\lambda_i) = \mathbf{w_i}'\boldsymbol{\gamma}$, where the logit function is $\text{logit}(p) = \log[p/(1-p)]$, and the inverse function $\text{logit}^{-1}(x) \equiv \text{expit}(x) \equiv \exp(x)/(1+\exp(x))$.

The reported value of y_{ij} is modeled as a Bernoulli random variable with probability π_{ij}, which is regressed on h fully observed response level predictors $\mathbf{x_{ij}} = (x_{ij1}, \ldots, x_{ijh})'$ also using a logit link. A mixture model allows the regression coefficients to differ depending on the value of z_i. If $z_i = 1$ then $\text{logit}(\pi_{ij}) = \mathbf{x_{ij}}'\boldsymbol{\alpha}$, else when $z_i = 0, \text{logit}(\pi_{ij}) = \mathbf{x_{ij}}'\boldsymbol{\beta}$. Both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of length l. The full hierarchical Bayesian model is

$$\begin{aligned} z_i|\lambda_i &\sim \text{Bernoulli}(\lambda_i) \\ \text{logit}(\lambda_i) &= \mathbf{w}_i'\boldsymbol{\gamma} \\ y_{ij}|\pi_{ij} &\sim \text{Bernoulli}(\pi_{ij}) \\ \text{logit}(\pi_{ij}) &= \mathbf{x_{ij}}'\boldsymbol{\alpha} z_i + \mathbf{x_{ij}}'\boldsymbol{\beta}(1-z_i), \end{aligned} \tag{4.1}$$

for $i = 1, \ldots, n,$ and $j = 1, \ldots, m_i$. Independent multivariate normal priors $p(\boldsymbol{\alpha}), p(\boldsymbol{\beta})$, and $p(\boldsymbol{\gamma})$ with known mean and variance parameters are assigned to the regression coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ respectively. Since label switching can be a problem in mixture models, the prior distributions on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are truncated such that the intercept coefficients $\alpha_1 \geq 0$ and $\beta_1 \leq 0$

$$\begin{aligned} \boldsymbol{\alpha} &\sim \mathcal{N}_p(\mathbf{m}_\alpha, \mathbf{V}_\alpha)\mathbf{I}\{\alpha_1 \geq 0\} \\ \boldsymbol{\beta} &\sim \mathcal{N}_p(\mathbf{m}_\beta, \mathbf{V}_\beta)\mathbf{I}\{\beta_1 \leq 0\} \\ \boldsymbol{\gamma} &\sim \mathcal{N}_q(\mathbf{m}_\gamma, \mathbf{V}_\gamma). \end{aligned} \tag{4.2}$$

Bayesian inference about the latent variable z_i and unknown parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ is made via sampling from the posterior densities of the unknown parameters conditional on the data.

4.1.1 Full Conditional Posterior Density Calculations

The form of the full joint posterior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, z_i, y_{ij} | \mathbf{x_{ij}}, \mathbf{w_i})$ is non-trivial and cannot be directly sampled from. Instead, posterior samples of the latent variable z_i and unknown regression coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are sampled from their full conditional posterior densities using an MCMC sampling algorithm. The full joint posterior distribution can be decomposed into the full conditionals

$$\begin{aligned}
&p(\boldsymbol{\alpha}|\mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}),\\
&p(\boldsymbol{\beta}|\mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\gamma}),\\
&p(z_i|\mathbf{x_{ij}}, \mathbf{w_i}, \mathbf{y_i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}), \qquad i = 1, \ldots, n,\\
&p(\boldsymbol{\gamma}|\mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}),
\end{aligned}$$

where $\mathbf{y_i} = (y_{i1}, \ldots, y_{im_i})'$ is the vector of m_i responses for subject i with $\mathbf{Y} = (\mathbf{y_1}, \ldots, \mathbf{y_n})'$. The other variables are defined as $\mathbf{X} = (\mathbf{x_{11}}, \ldots, \mathbf{x_{nm_i}})'$ is the N x l matrix of known covariates for all $N = \sum_i m_i$ observations, $\mathbf{Z} = (z_1, \ldots, z_n)'$ is the n-vector of underlying true values for all n subjects, and $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)'$ is the n x q matrix of predictors for z_i. The probability density function for $\boldsymbol{\theta} \sim \mathcal{N}_p(\mathbf{m}_\theta, \mathbf{V}_\theta)$ has the form

$$p(\boldsymbol{\theta}) = (2\pi)^{-\frac{p}{2}} |\mathbf{V}_\theta|^{-\frac{1}{2}} \exp\big[-\frac{1}{2}(\boldsymbol{\theta} - \mathbf{m}_\theta)' \mathbf{V}_\theta^{-1} (\boldsymbol{\theta} - \mathbf{m}_\theta)\big].$$

Taking a log and dropping constants results in a functional form

$$\log p(\boldsymbol{\theta}) \propto -\frac{1}{2}(\boldsymbol{\theta} - \mathbf{m}_\theta)' \mathbf{V}_\theta^{-1} (\boldsymbol{\theta} - \mathbf{m}_\theta). \tag{4.3}$$

This form is used in further posterior density calculations.

Derivation of $p(\boldsymbol{\alpha}|\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ and $p(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}, \mathbf{Y})$. Since $\boldsymbol{\gamma}$ affects $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ indirectly through $\mathbf{Z}$, the conditional posterior densities

$$p(\boldsymbol{\alpha}|\mathbf{Y}, \mathbf{Z}, \mathbf{X}) \propto L(\mathbf{Y}|\boldsymbol{\alpha}, \mathbf{X}, \mathbf{Z}) p(\boldsymbol{\alpha})$$

and

$$p(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{Z}, \mathbf{X}) \propto L(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}) p(\boldsymbol{\beta})$$

do not directly depend on $\boldsymbol{\gamma}$. The sampling density of the response data y_{ij} is

$$f(y_{ij}|\pi_{ij}) = \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}, \qquad y_{ij} \in 0, 1 \tag{4.4}$$

for $i = 1, \ldots, n,\ j = 1, \ldots, m_i$. Let

$$\boldsymbol{\eta}_{ij} = \mathbf{x_{ij}}^{'}\boldsymbol{\alpha} z_i + \mathbf{x_{ij}}^{'}\boldsymbol{\beta}(1-z_i)$$

then

$$\pi_{ij} = \text{expit}(\boldsymbol{\eta}_{ij}).$$

Taking the log and dropping constants, equation (4.4) can be written as

$$\log f(y_{ij}|\boldsymbol{\eta}_{ij}) = y_{ij}\boldsymbol{\eta}_{ij} - \log[1 + \exp(\boldsymbol{\eta}_{ij})].$$

Specifically,

$$\log f(y_{ij}|\mathbf{x_{ij}}, \boldsymbol{\alpha}, z_i)|_{z_i=1} = y_{ij}\mathbf{x_{ij}}^{'}\boldsymbol{\alpha} - \log(1 + e^{\mathbf{x_{ij}}^{'}\boldsymbol{\alpha}}), \tag{4.5}$$

and

$$\log f(y_{ij}|\mathbf{x_{ij}}, \boldsymbol{\beta}, z_i)|_{z_i=0} = y_{ij}\mathbf{x_{ij}}^{'}\boldsymbol{\beta} - \log(1 + e^{\mathbf{x_{ij}}^{'}\boldsymbol{\beta}}). \tag{4.6}$$

The full log conditional posterior distributions for the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ then are

$$\begin{aligned}\log p(\boldsymbol{\alpha}|\mathbf{X}, \mathbf{Z}, \mathbf{Y}) &\propto \log L(\mathbf{Y}|\boldsymbol{\alpha}, \mathbf{X}, \mathbf{Z}) + \log p(\boldsymbol{\alpha}) \\ &\propto \sum_{ij|z_i=1} f(y_{ij}|\boldsymbol{\alpha}, \mathbf{x_{ij}}, z_i) - \frac{1}{2}(\boldsymbol{\alpha} - \mathbf{m}_\alpha)^{'}\mathbf{V}_\alpha^{-1}(\boldsymbol{\alpha} - \mathbf{m}_\alpha),\end{aligned} \tag{4.7}$$

and

$$\begin{aligned}\log p(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}, \mathbf{Y}) &\propto \log L(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{X}, \mathbf{Z}) + \log p(\boldsymbol{\beta}) \\ &\propto \sum_{ij|z_i=0} f(y_{ij}|\boldsymbol{\beta}, \mathbf{x_{ij}}, z_i) - \frac{1}{2}(\boldsymbol{\beta} - \mathbf{m}_\beta)^{'}\mathbf{V}_\beta^{-1}(\boldsymbol{\beta} - \mathbf{m}_\beta).\end{aligned} \tag{4.8}$$

Derivation of $p(z_i|\mathbf{x_{ij}}, \mathbf{w_i}, \mathbf{y_i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$. The conditional sampling density of $z_i|\lambda_i$ is

$$f(z_i|\lambda_i) = \lambda_i^{z_i}(1-\lambda_i)^{1-z_i} \qquad z_i \in 0, 1, \tag{4.9}$$

for $i = 1, \ldots, n$. Letting $\lambda_i = \text{expit}(\mathbf{w_i}'\boldsymbol{\gamma})$ and taking the log, this can be written as

$$\log f(z_i|\boldsymbol{\gamma}, \mathbf{w_i}) = z_i \mathbf{w_i}'\boldsymbol{\gamma} - \log(1 + e^{\mathbf{w_i}'\boldsymbol{\gamma}}), \tag{4.10}$$

for $i = 1, \ldots, n$. The conditional posterior of the underlying variable z_i depends on all other parameters and data from subject i. Let k_{1i} and k_{0i} be the log-posterior of z_i evaluated at $z_i = 1$ and $z_i = 0$ respectively

$$k_{1i} = \log f(z_i|\boldsymbol{\gamma}, \mathbf{w_i})|_{z_i=1} + \sum_j \log f(y_{ij}|\boldsymbol{\alpha}, \mathbf{x_{ij}}, z_i)|_{z_i=1}, \tag{4.11}$$

$$k_{0i} = \log f(z_i|\boldsymbol{\gamma}, \mathbf{w_i})|_{z_i=0} + \sum_j \log f(y_{ij}|\boldsymbol{\beta}, \mathbf{x_{ij}}, z_i)|_{z_i=0}. \tag{4.12}$$

The full conditional posterior distribution of z_i for $i = 1, \ldots, n$ is Bernoulli

$$z_i|\mathbf{w_i}, \mathbf{x_{ij}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, y_{ij} \sim \text{Bernoulli}\Big(\frac{\exp(k_{1i})}{\exp(k_{1i}) + \exp(k_{0i})}\Big). \tag{4.13}$$

Derivation of $p(\boldsymbol{\gamma}|\mathbf{Z}, \mathbf{W})$. The log full conditional posterior density of the unknown vector $\boldsymbol{\gamma}$ depends on $\mathbf{Z}$ and $\mathbf{W}$

$$p(\boldsymbol{\gamma}|\mathbf{Z}, \mathbf{W}) \propto L(\mathbf{Z}|\boldsymbol{\gamma}, \mathbf{W})p(\boldsymbol{\gamma}).$$

Taking logs and dropping constants this can be written as

$$\begin{aligned} \log p(\boldsymbol{\gamma}|\mathbf{Z}, \mathbf{W}) &\propto \log L(\mathbf{Z}|\boldsymbol{\gamma}, \mathbf{W}) + \log p(\boldsymbol{\gamma}) \\ &\propto \sum_i \log f(z_i|\boldsymbol{\gamma}, \mathbf{w_i}) - \frac{1}{2}(\boldsymbol{\gamma} - \mathbf{m}_\gamma)'\mathbf{V}_\gamma^{-1}(\boldsymbol{\gamma} - \mathbf{m}_\gamma). \end{aligned} \tag{4.14}$$

All log full conditional posterior densities have now been defined. Next I discuss how the posterior samples are generated.

4.1.2 Sampling Algorithm

To sample from the posteriors $p(\boldsymbol{\alpha}|\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, $p(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, and $p(\boldsymbol{\gamma}|\mathbf{Z}, \mathbf{W})$, I use an adaptive random walk Metropolis-within-Gibbs sampling algorithm with multivariate normal proposal

distributions. A Gibbs step is used to sample z_i directly from equation (4.13). The sampling order is $\boldsymbol{\gamma}$, $z_{i,i=1,\ldots,n}$, $\boldsymbol{\alpha}$ and then $\boldsymbol{\beta}$. To reduce computational burden during simulation, the acceptance probability from equation (2.12) is re-written as

$$\rho_{\boldsymbol{\theta}} = \min\left[1, \exp\left(\log L(\mathbf{y}|\boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta}^*) - \log L(\mathbf{y}|\boldsymbol{\theta}^{(\ell-1)}) + \log p(\boldsymbol{\theta}^{(\ell-1)})\right)\right], \tag{4.15}$$

where $\boldsymbol{\theta} = \boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ each individually and $\boldsymbol{\theta}^*$ is the candidate value. To incorporate the constraints placed on the intercepts α_1 and $\beta_1, p(\boldsymbol{\alpha}^*) = 0$ if $\mathbf{I}\{\alpha_1 < 0\}$ and $p(\boldsymbol{\beta}^*) = 0$ if $\mathbf{I}\{\beta_1 > 0\}$. Each of the four components of $\rho_{\boldsymbol{\theta}}$ are calculated once then updated as necessary. This means that if the acceptance criteria $u \sim U(0,1) < \rho_{\boldsymbol{\theta}}$ is met, set

$$\boldsymbol{\theta}^{(\ell)} = \boldsymbol{\theta}^*, \tag{4.16a}$$

$$\log L(\mathbf{Y}|\boldsymbol{\theta}^{(\ell)}) = \log L(\mathbf{Y}|\boldsymbol{\theta}^*), \text{ and} \tag{4.16b}$$

$$\log p(\boldsymbol{\theta}^{(\ell)}) = \log p(\boldsymbol{\theta}^*). \tag{4.16c}$$

Otherwise retain the previous values

$$\boldsymbol{\theta}^{(\ell)} = \boldsymbol{\theta}^{(\ell-1)}, \tag{4.17a}$$

$$\log L(\mathbf{y}|\boldsymbol{\theta}^{(\ell)}) = \log L(\mathbf{y}|\boldsymbol{\theta}^{(\ell-1)}), \text{ and} \tag{4.17b}$$

$$\log p(\boldsymbol{\theta}^{(\ell)}) = \log p(\boldsymbol{\theta}^{(\ell-1)}). \tag{4.17c}$$

Lastly, define the average probability that a student is female as

$$\bar{\lambda}^{(\ell)} = \frac{1}{n}\sum_{i=1}^{n} \lambda_i^{(\ell)}, \tag{4.18}$$

where $\lambda^{(\ell)}$ is the ℓ^{th} sampled value for λ_i. The value of $\bar{\lambda}^{(\ell)}$ then is the average probability of $z = 1$ at iteration ℓ.

4.1.3 Editing and Imputation

This section discusses the process of jointly editing the inconsistent reports and imputing the missing data in a stochastic fashion. Introduce Z_i, student i's true value of the characteristic. To correctly account for the error generated in these processes, $Z_i^{(m)}$, $m = 1, \ldots, M$ imputed

and edited vectors are created. The data sets containing these updated vectors are called *MEMI's* (Multiple Edit Multiple Imputation) (Ghosh-Dastidar and Schafer, 2003). I propose two variations on a joint imputation and editing process for repeated measures with missing or inconsistent data.

The first editing and imputation procedure is a *full edit*, where all records are edited regardless of whether an IRM was observed. Using the editing indicator notation where $E_i = 1$ if an IRM was observed and 0 otherwise, this would set $E_i = 1$ for all $i = 1, \ldots, n$ students. The second type of edit is a *limited edit* which only changes records that are observed to be inconsistent.

Full edit. There is a non-zero probability that a student could consistently mis-report the characteristic in question on all surveys. This acknowledges that someone with $Z_i = 0$ could report $y_{ij} = 1$ all m_i times they took the survey, or report 1 twice and not respond to the question the last time. This would argue that everyone in the sample should be edited ($E_i = 1$ for all n), not just those observed to report inconsistently. This method sets $Z_i^{(\ell)} = z_i^{(\ell)}$ for all $i = 1, \ldots, n$, which is equivalent to retaining the drawn values from the Gibbs sampling step in equation (4.13).

Limited edit. Under the limited edit only participants who provide multiple years of data are candidates for the limited editing, a single response such as 0 or 1 is not edited. To implement the limited edit, Z_i is set as the reported value y_{i1} from the first reported survey if the student consistently reported across all surveys ($E_i = 0$). For students with inconsistent reports, ($E_i = 1$), a draw from the posterior sample of z_i is used for Z_i.

Imputation If the student did not report the characteristic of interest on any survey they participated in, the missing data indicator M_i is equal to 1 for student i and 0 otherwise. For students with $M_i = 1, Z_i^{(\ell)}$ is drawn as a random Bernoulli variable with probability $\text{expit}\{\mathbf{w}_i'\boldsymbol{\gamma}^{(\ell)}\}$ if full covariate data was provided, and $\sum_{t=1}^{\tilde{n}} \bar{\lambda}^{(\ell)}$ if not.

In summary the full edit is

$$Z_i^{(\ell)} \quad \begin{cases} = z_i^{(\ell)} \text{ if } M_i = 0 \\ \sim \text{Bernoulli}(\text{expit}\{\mathbf{w}_i' \boldsymbol{\gamma}^{(\ell)}\}) \\ \qquad \text{for all } y_{ij} \text{ missing } (M_i = 1) \text{ and } \mathbf{w}_i \text{ observed} \\ \sim \text{Bernoulli}(\sum_{ell=1}^{\tilde{n}} \bar{\lambda}^{(\ell)}) \text{ if } (M_i = 1) \\ \qquad \text{and not all covariate information is available} \end{cases}, \tag{4.19}$$

and the limited edit is

$$Z_i^{(\ell)} \quad \begin{cases} = y_{ij} \text{ for consistent reports } (E_i = 0) \\ = z_i^{(\ell)} \text{ for inconsistent reports} (E_i = 1) \\ \sim \text{Bernoulli}(\text{expit}\{\mathbf{w}_i' \boldsymbol{\gamma}^{(\ell)}\}) \\ \qquad \text{for all } y_{ij} \text{ missing } (M_i = 1) \text{ and } \mathbf{w}_i \text{ observed} \\ \sim \text{Bernoulli}(\sum_{\ell=1}^{\tilde{n}} \bar{\lambda}^{(\ell)}) \text{ if } (M_i = 1) \\ \qquad \text{and not all covariate information is available.} \end{cases} \tag{4.20}$$

I present two examples of IRM's, one editing gender and one editing birthplace.

4.2 Example 1: Analysis of Gender

Gender is one of the most important covariates in behavioral intervention studies as interventions can have different effects on males and females (DeRosa et al., 2012). Inconsistent responses on gender across surveys for the same student could confound how gender modifies the effect of an intervention on a particular outcome. A value of gender can be determined if gender is asked only once, and that single value is used by the analyst for all time points, provided everyone can be assumed to answer truthfully. The Project Connect study asked participants their gender on each survey without predetermined response verification. This opened the door to inconsistent reporting, but also to the possibility of fixing the errors through statistical analysis. I apply the IRM model to the repeated self-reported gender to

model each student's true gender, but first explore the probability of a student being female given their response pattern using a simpler probability model with no covariates.

Since students participated in 1 to 4 surveys, there are 30 distinct response patterns of Male (M) and Female (F). For students who participated all 4 years there are 16 possible response patterns: MMMM, MMMF, MMFM, MFMM, FMMM, MMFF, MFMF, MFFM, FMMF, FMFM, FFMM, MFFF, FMFF, FFMF, FFFM, and FFFF. Students with 3 surveys have 8 response patterns: MMM, MMF, MFM, FMM, MFF, FMF, FFM, FFF, and students with 2 surveys have 4 response patterns: MM, MF, FM, and FF.

Consider a student who participates for all four years. If they report being male on all four surveys, then one could believe this student truly is a male. Similarly for an FFFF response pattern it is extremely likely this person is female. But what about someone with only three years of data? Does MMF have the same strength as MMMF in determining the underlying true gender of the student in the presence of conflicting responses? Most likely this person is male, but there is a small chance they are female. Depending on the probability of a student mis-reporting their gender, the probability they are female varies from one response pattern to another, especially if the probability of mis-reporting differs by gender.

Let $p(F)$ be the prior probability of a student being female where $p(M) = 1 - p(F)$ is the probability of being male. Let π_M and π_F be the mis-reporting probability for males and females respectively. Assuming reports are independent across surveys given π_M or π_F, and using Bayes' formula, the probability of a student being female given a report of n_m M's and n_f F's is

$$p(F|\text{data}) = \frac{p(F)\pi_F^{n_m}(1-\pi_F)^{n_f}}{p(F)\pi_F^{n_m}(1-\pi_F)^{n_f} + p(M)\pi_M^{n_f}(1-\pi_M)^{n_m}}. \tag{4.21}$$

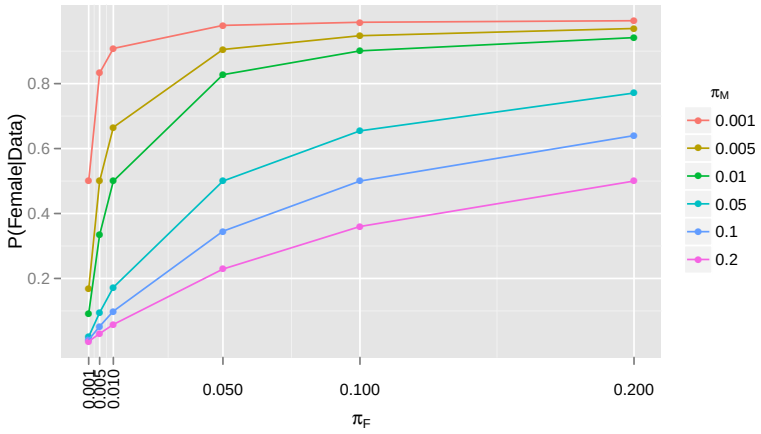
Consider a student who reported being male once, and female on each of 1 to 3 additional surveys: MF, MFF, and MFFF. Using equation (4.21), the probability this student is female using varying levels of π_M and π_F under each of the three possible IRM cases are plotted in Figure 4.1.

Figure 4.1(a) shows that when the probability that female mis-reports their gender is high

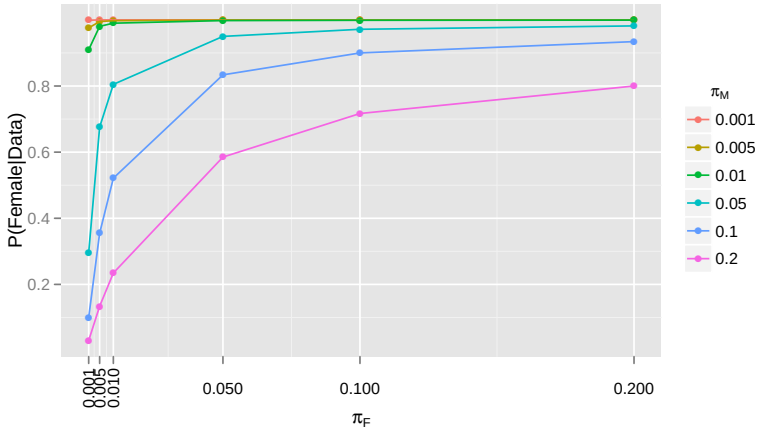
$\pi_F = .2$ and the probability of a male mis-reporting their gender is low $\pi_M = .001$ the chance that a student reporting MF is Female is .994. This is intuitive in that if a male is less likely to lie than a female, and we observe an inconsistent report, there is a higher probability that the report came from a female. If males and females are equally likely to lie, and there is only one report on each, then it is equally likely a report of MF came from a female as from a male.

Figures 4.1(b) and 4.1(c) show that an increase in reporting consistency however plays a much larger role than the probability of mis-reporting. Figure 4.1(c) specifically shows that even if both females and males lie 20% of the time there is more than a .9 probability that a report of MFFF came from a female. Only when $\pi_F < .005$ and $\pi_M = .2$ does the probability that this report comes from a female drops below .5 given the values for π_M and π_F.

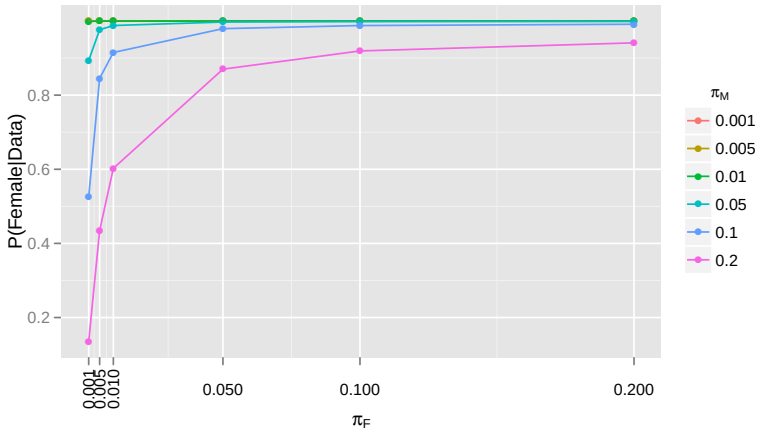
This probability model demonstrates the roles that the number of consistent reports and probability of mis-reporting play in the overall posterior probability of the student being female. However, this model does not include information on gender or the reporting of gender that can be found in other covariates measured on the student. To properly account for the amount and consistency of the self-report data on gender using information contained in other variables, the IRM model can be used to model students' true gender and perform imputation and editing in a stochastic manner.



(a) MF



(b) MFF



(c) MFFF

Figure 4.1: Probability of being female given for various values of π_M and π_F.

4.2.1 Data

The data used for this example comes from $N = 36,327$ observations on $n = 26,606$ students who are not missing data on selected predictors. I use subject level binary indicator variables for carrying weapons (WEAP_i) and fighting (FIGHT_i) as predictors in the model for the probability $\lambda_i = p(z_i = 1)$ that student i is a female. As these are variables that are asked each year, as a practical matter I have aggregated the responses across surveys by setting $\text{WEAP}_i = 1$ if student i reported that they carried a weapon at least one day (in the 30 days prior) on any survey, and $\text{FIGHT}_i = 1$ if student i reported engaging in at least one fight in the past year on any survey. Since this becomes an indicator of ever fighting or ever carrying a weapon (within the specified time frames), non-responses are imputed as 0. While this deterministic imputation may seem to be contrary to the methodology presented in this dissertation, it is performed here to simplify the example and it is believed to be fairly accurate. Weapons, fighting plus an intercept make up the subject level covariates $\mathbf{w_i} = (1, \text{WEAP}_i, \text{FIGHT}_i)'$.

I use one response level covariate, AGE_{ij}, to model the probability $\pi_{ij} = P(y_{ij} = 1)$ of a student reporting being female given their true gender. The predictor AGE_{ij} is centered by subtracting the mean and standardized by dividing by the standard deviation calculated across all observations. This variable plus an intercept make up the response level covariates $x_{ij} = (1, \text{AGE}_{ij})'$.

4.2.2 Prior Distributions and Simulation Settings

Vague but proper normal priors are placed on all regression coefficients. The prior means for the reporting model $\mathbf{m}_\alpha$ and $\mathbf{m}_\beta$ are set at $(5, 0)'$ and $(-5, 0)'$ respectively. This reflects the prior belief that there is a $\text{expit}(5) = 0.993$ probability a student will correctly report their gender, and that a priori age has no expected effect on how a student reports their gender. The prior mean for the true gender model $\mathbf{m}_\gamma = (0, -1, -0.5)'$ is set to reflect the prior belief that females have a lower likelihood of carrying weapons or fighting (Centers for Disease Control and Prevention, 2010).

The prior variances for the regression coefficients are set in a fashion similar to the prior variances for Zellner g-priors (Zellner, 1983). The prior variances are

$$\mathbf{v}_\alpha = \mathbf{v}_\beta = n_0^{-1} * N(X^{'}X)^{-1} \tag{4.22a}$$

and

$$\mathbf{v}_\gamma = n_1^{-1} * n(W^{'}W)^{-1}, \tag{4.22b}$$

where N is the total number of observations, n the total number of unique students, and n_0 and n_1 are scaling factors. These scaling factors determine the strength of the prior; the larger the factor the stronger the prior information is. For example $n_0 = 100$ is equivalent to adding 100 additional observations worth of information, creating an informative prior distribution on the response level model. The prior covariance matrices are

$$\mathbf{v}_\alpha = \mathbf{v}_\beta = \begin{array}{c} \begin{array}{cc} \text{Intercept} & \text{AGE} \end{array} \\ \left(\begin{array}{cc} 0.2000 & -0.0004 \\ & 0.2004 \end{array} \right) \end{array}, \tag{4.23}$$

and

$$\mathbf{v}_\gamma = \begin{array}{c} \begin{array}{ccc} \text{Intercept} & \text{WEAP} & \text{FIGHT} \end{array} \\ \left(\begin{array}{ccc} 0.603 & 0.976 & 0.071 \\ & 2.690 & -0.551 \\ & & 1.104 \end{array} \right) \end{array}, \tag{4.24}$$

where $n_0 = n_1 = 5$.

Simulation settings. The model for inconsistent repeated measures of gender is fit with Markov chain Monte Carlo (MCMC) simulation techniques using $s = 5$ parallel chains. Phase 1 of each chain consists of $D = 2$ blocks of $m = 1,000$ iterations each and is discarded as the burn-in. Phase 2 simulation is run for $M_2 = 40,000$ additional iterations per chain, retaining every $k = 40^{\text{th}}$ iteration, resulting in a final sample size of $\tilde{n} = 5,000$.

4.2.3 Results

Convergence Diagnostics. Figure 4.2 shows trace plots on the left and density plots on the right for samples generated in Phase 2 from each of the 5 chains. The trace plots demonstrate the adequate mixing and convergence of the multiple chains. The density plots also indicate that the chains have converged to the same smooth distribution. The thick grey line plots the prior distribution for that parameter. Figure 4.3 displays the cumulative acceptance rates for each vector of regression coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ per chain. The acceptance rate per chain is calculated per block as the number of times the candidate value was accepted up to iteration ℓ divided ℓ. The final acceptance rate for a parameter vector is calculated as the average value across chains at iteration ℓ and displayed in the upper right corner of Figure 4.3 and are 0.21, 0.30, and 0.18 for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ respectively. The autocorrelation plots on the left in Figure 4.4 indicate that using a thin of $k = 40$ resulted in an approximately uncorrelated posterior samples, and the Gelman-Rubin-Brooks diagnostic plots on the right show that for all regression parameters the shrink factor is close to 1. This is also a good indication of convergence.

Posterior Summary. Table 4.1 gives summary statistics for the Inconsistent Repeated Measures gender model including posterior means and standard deviations for the regression coefficients and the average probability that a student is female. To aid interpretation, regression coefficients have been transformed into Odds Ratios (OR= $\exp(\theta)$), with intercepts transformed into probabilities (p=logit$^{-1}(\theta)$) and displayed in *italics*, both with corresponding 95% posterior intervals.

A student who reported carrying a weapon in the 30 days prior to the survey date on any survey has .38 (95%PI .35, .42) lower odds of being female than one not carrying a weapon during that time frame. Students reporting fighting at least once in the year prior to the survey date have .76 (.72, .81) lower odds of being female than a student who did not engage in any fighting. The average probability a student in this sample is female is .544 (.543, .546). The probability of reporting being female is .995 (.993, .997) for females and .003 (.002, .005) for males at the sample mean age of 14.9. For both genders, as age increases

Parameter	Interpretation	Mean	SD	OR/p	2.5%	97.5%	$p(\theta > 0)$
$\bar{\lambda}$	p(Female)	0.544	0.0007	-	0.543	0.546	-
γ_1	Intercept	−0.279	0.0230	*0.431*	*0.420*	*0.442*	<0.001
γ_2	Carried a weapon 30d	−0.964	0.0489	0.381	0.347	0.420	<0.001
γ_3	Fought past 12mo	−0.275	0.0293	0.760	0.718	0.805	<0.001
Female							
α_1	Intercept	5.323	0.1738	*0.995*	*0.993*	*0.997*	1.000
α_2	Age (standardized)	−0.347	0.2001	0.707	0.475	1.041	0.038
Male							
β_1	Intercept	−5.816	0.2380	*0.003*	*0.002*	*0.005*	<0.001
β_2	Age (standardized)	−0.329	0.2103	0.720	0.479	1.083	0.058

Table 4.1: Summary of the posterior distribution for the gender model parameters. Odds Ratios (OR) or probabilities (p in *italics*) with corresponding 95% posterior intervals are included.

the odds of reporting being female decreases. Figure 4.5 plots the probability that a student will correctly report their gender decreases as age increases for females, but increases as age increases for males. The lines are calculated as

$$p(a) = \text{expit}(5.3225 - .3470 * a),$$
$$p(a) = 1 - \text{expit}(-5.8157 - .3288 * a)$$

for females and males respectively, where a is the standardized age. How these posterior parameter estimates are used in an editing and imputation procedure is discussed next.

MEMI Results. Using equation (4.20), data are edited and imputed $M = 20$ times. To confirm the editing worked in a manner consistent with logic and comparable to the deterministic editing rules, Table 4.2 displays the count of males and females in the data set after editing and imputing, ordered by the number of surveys the student took m_i and the number of times they reported being female $\#F$. Results are displayed for limited edit MEMI #1, #7, and #20, and full edit #3, #9, and #10 to illustrate the variation in editing

results. The last two columns display counts of males and females averaged across all 20 MEMI's separately for the limited and full edit models.

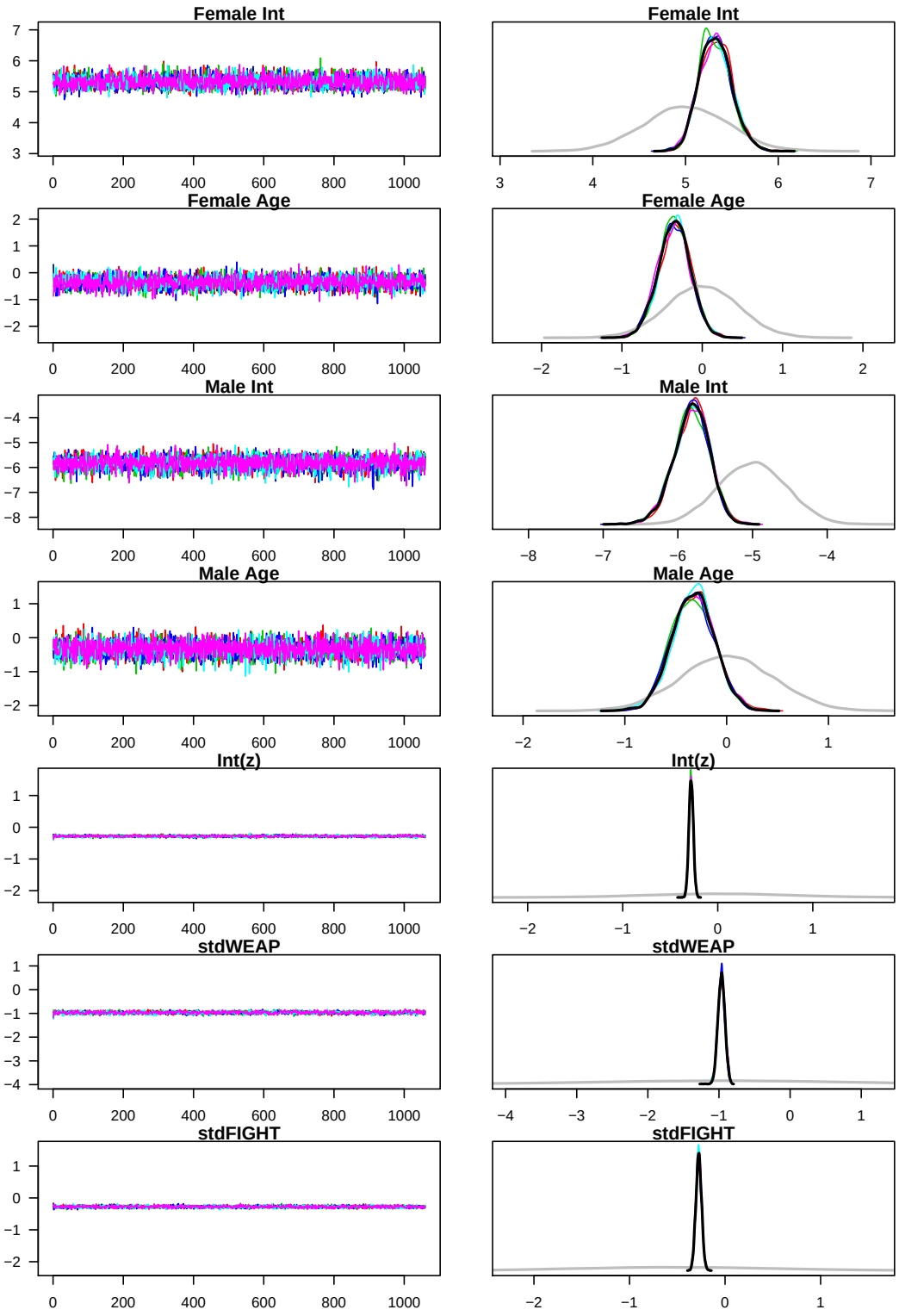


Figure 4.2: Trace (left) and posterior density (right) plots for the gender regression parameters in the IRM model. Prior densities are drawn in grey, each of the 5 chains has its own color with the average density drawn with the thick black line.

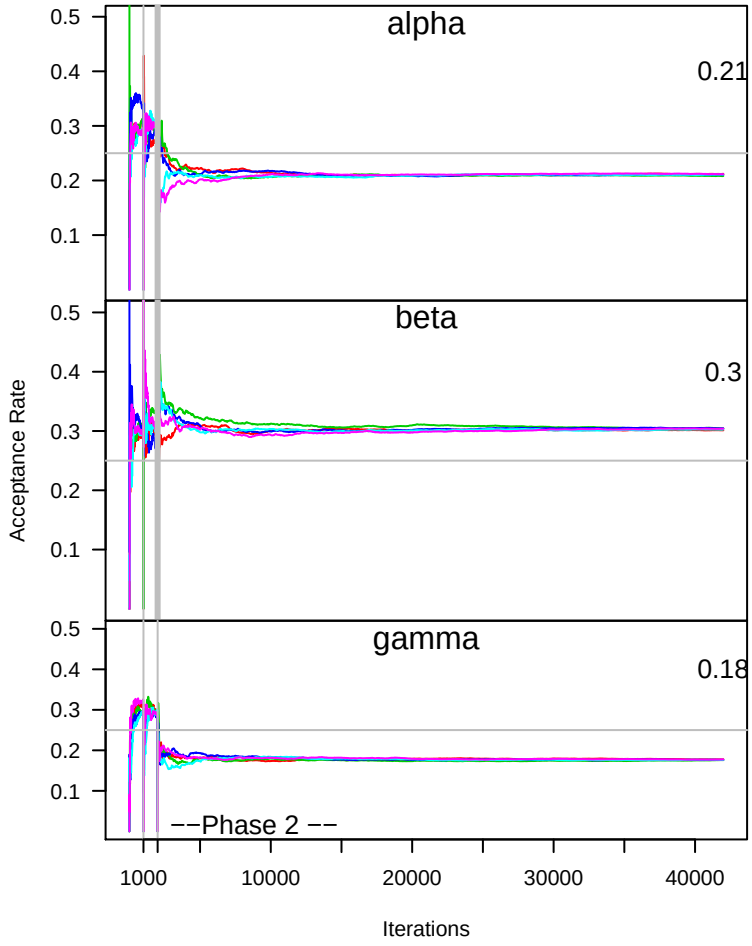


Figure 4.3: Cumulative acceptance rates for the regression coefficients. Grey vertical bars mark where the proposal variance is re-estimated.

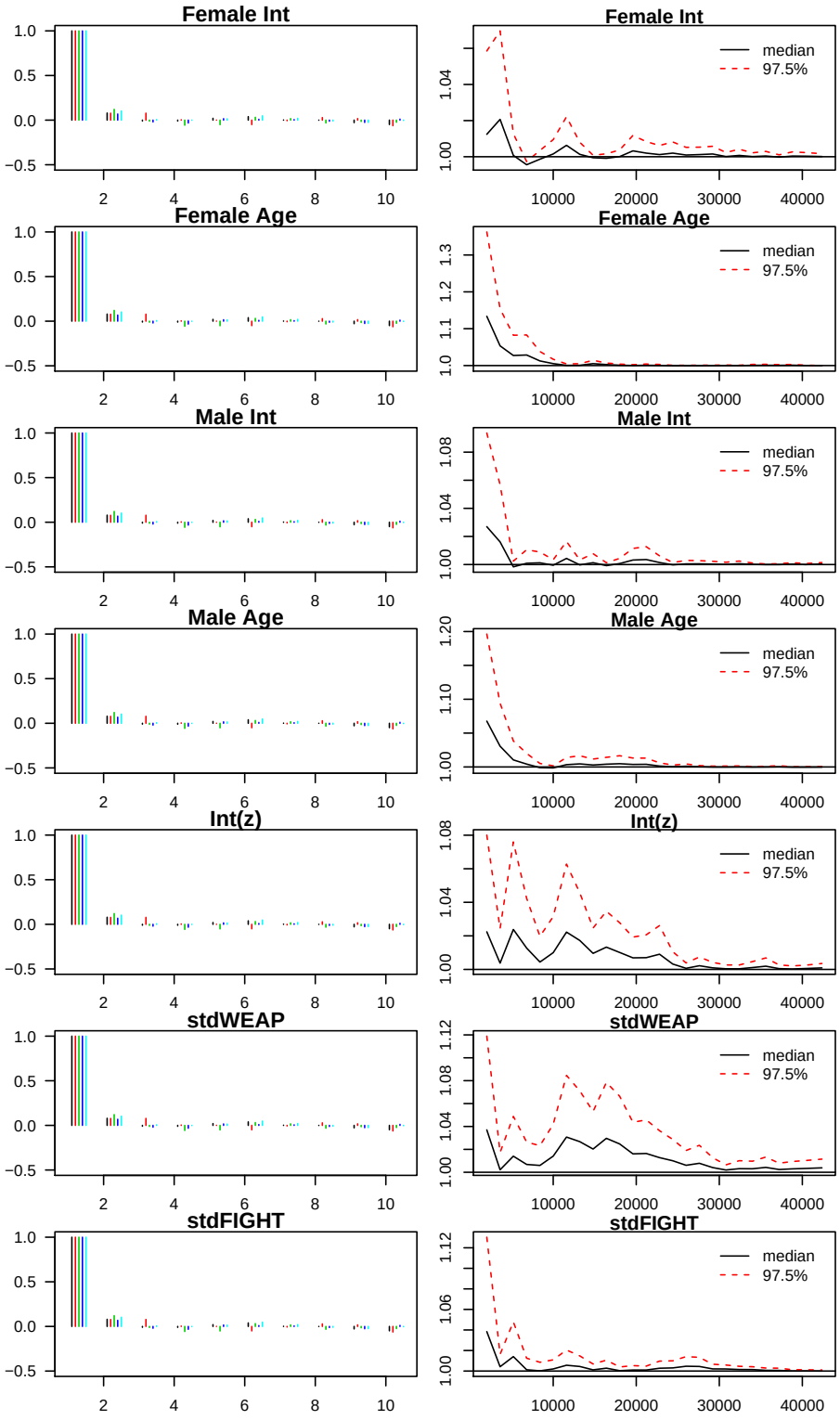


Figure 4.4: Autocorrelation (left) after a thin of 40 and and Gelman-Rubin (right) diagnostic plots for the gender regression parameters in the IRM model. The desirable target for the Gelman-Rubin diagnostic is 1.

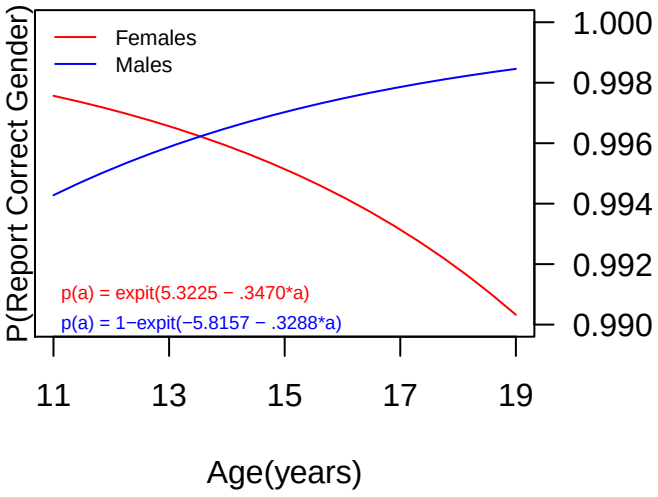


Figure 4.5: Probability of correctly reporting gender as a function of age for females and males.

m_i	$\#F$	Limited Edit						Full Edit						Average			
		MEMI #1		MEMI #7		MEMI #20		MEMI #3		MEMI #9		MEMI #10		Limited		Full	
		M	F	M	F	M	F	M	F	M	F	M	F	M	F	M	F
0	0	**108**	**137**	**117**	**128**	**114**	**131**	**112**	**133**	**104**	**141**	**118**	**127**	**111.7**	**133.4**	**111.65**	**133.35**
1	0	9544	0	9544	0	9544	0	9497	**47**	9488	**56**	9434	**110**	9544	0	**9474.15**	**69.85**
1	1	0	11065	0	11065	0	11065	**41**	11024	**48**	11017	**21**	11044	0	11065	**31.3**	**11033.7**
2	0	1391	0	1391	0	1391	0	1391	0	1391	0	1389	**2**	1391	0	**1390.75**	**0.25**
2	1	**20**	**44**	**11**	**53**	**6**	**58**	**9**	**55**	**12**	**52**	**7**	**57**	**9.75**	**54.25**	**9.85**	**54.15**
2	2	0	1744	0	1744	0	1744	**1**	1743	0	1744	0	1744	0	1744	**0.15**	**1743.85**
3	0	813	0	813	0	813	0	813	0	813	0	813	0	813	0	813	0
3	1	**7**	0	**7**	0	**7**	0	**7**	0	**7**	0	**6**	**1**	**6.95**	**0.05**	**6.95**	**0.05**
3	2	**1**	**40**	0	**41**	0	**41**	0	**41**	**1**	**40**	0	**41**	**0.1**	**40.9**	**0.1**	**40.9**
3	3	0	1011	0	1011	0	1011	0	1011	0	1011	0	1011	0	1011	0	1011
4	0	398	0	398	0	398	0	398	0	398	0	398	0	398	0	398	0
4	1	**2**	0	**2**	0	**2**	0	**2**	0	**2**	0	**2**	0	**2**	0	**2**	0
4	2	0	**1**	0	**1**	0	**1**	0	**1**	0	**1**	0	**1**	0	**1**	0	**1**
4	3	0	**24**	0	**24**	0	**24**	0	**24**	0	**24**	0	**24**	0	**24**	0	**24**
4	4	0	501	0	501	0	501	0	501	0	501	0	501	0	501	0	501

Table 4.2: Count of males and females from 3 different MEMI data sets under each the limited and full editing procedure. Average columns are calculated to show the cell counts averaged across all 20 MEMI data sets. Rows are organized by the number of surveys taken m_i and the number of times a gender of female was reported #F. Imputed values are **blue**, edited values are **green** and edits performed on records with no observed inconsistencies are shown in **orange**.

Out of the 20,609 students with only one survey, the full edit changed the gender of 101 ($< 1\%$) students, on average. Of the 3,199 students with exactly 2 surveys, 64 (2%) students reported being male on one survey and female on the other. Of these 64, on average 54.14 (84.5%) were edited as female. Across all 20 MEMI's under the full edit procedure, at most 2 students consistently reporting being male on both surveys were edited to be Female, and 1 student with 2 consistent reports of female was edited to Male. For students providing data on 3 surveys, the results of the limited editing procedure are almost completely consistent with a majority rules deterministic edit. MEMI #1 and #9 edited one person who reported being female 2 times out of 3 to be male, and MEMI #10 edited one person to be female who reported being female on only 1 out of 3 surveys. For students who provided 4 surveys there were no tied inconsistent responses of MMFF, and the results of both editing procedures were consistent with the majority rules deterministic edit.

Table 4.3 shows that the percent of females ranges from 53.68% to 54.88% across MEMI's for the limited edit, and 53.80% to 54.04% for the full edit. These results are combined using Rubin's rules and 95% intervals are created that reflect the variation between the edits. Rubin's rules to combine estimates across imputations are given in equation (3.1). The average across all imputations for %F is $Q = .5428$. The within-imputation variance is the average of the individual variances, $U = 9.2\times10^{-6}$ and the between-imputation variance is the variance of the estimates, $B = 6.0\times10^{-8}$. The total variance is the weighted average of the between and within variances $T = 9.3\times10^{-6}$ and the 95% interval for the percent female that accounts for the multiple imputations and multiple edits is $Q \pm 1.96 * \sqrt{T} = (.5368, .5488)$.

	Limited Edit Model		Full Edit Model	
MEMI #	F	%F	F	%F
1	14571	54.27%	14580	54.30
2	14571	54.27%	14614	54.43
3	14567	54.25%	14572	54.27
4	14574	54.28%	14595	54.36
5	14580	54.30%	14624	54.46
6	14565	54.24 %	14617	54.44
7	14567	54.25 %	14598	54.37
8	14567	54.25 %	14616	54.43
9	14580	54.30 %	14588	54.33
10	14587	54.33 %	14678	54.66
11	14568	54.25 %	14620	54.45
12	14576	54.28 %	14606	54.40
13	14567	54.25 %	14612	54.42
14	14585	54.32 %	14614	54.43
15	14578	54.29 %	14621	54.45
16	14582	54.31 %	14617	54.44
17	14575	54.28 %	14613	54.42
18	14571	54.27 %	14611	54.42
19	14572	54.27 %	14615	54.43
20	14580	54.30%	14643	54.53
Combined	54.28 (53.68%, 54.88%)		54.42 (53.80%, 55.04%)	

Table 4.3: Number and percent female (N=26,606) after $M = 20$ edits under each the limited and full editing models. The mean and 95% intervals for the percent female after combining across edits using Rubin's rules is shown in the last row.

4.2.4 Discussion

Sample Size and Missing Data Out of the 26,851 students in the data set, 245 (0.9%) did not report a gender on any survey they took and 57 (0.2%) students report different genders in different years. Less than a 1% error rate is not normally of concern to the researchers, but examining gender is a natural starting place. It is a single variable that has been reported repeatedly over time with occasional response error and it is a variable that is relatively easy for researchers and analysts to understand.

Figure 4.6 shows the results from different ways of calculating the %Female. The most naive way is shown at the bottom in khaki. This calculates the proportion of females over all reported genders including repeated measures while ignoring missing data. Moving up to the dark goldenrod we only include students in the longitudinal cohort and calculate the naive proportion of all reported genders. This does not account for repeated measures nor missing data. Both the pale blue and slate blue bars are on the subject level. Gender at this level has been subject to the PC editing and imputation rules. The olive green bars at the top are a result of the limited (light) and full (dark) multiple editing procedure. Credible intervals instead of point estimates for the proportion of females are displayed to reflect the uncertainty in the editing.

The patterns in Table 4.2 indicate that for this example, a simple majority rules deterministic edit could be suitable for people with an odd number of surveys. For those with an even number of surveys where no majority can be determined, implementing a multiple edit procedure is better than a static gender assignment. The number of students consistently reporting their gender on two surveys that were subsequently edited under the full edit procedure is fairly large and might be unacceptable to researchers. A middle ground between the limited and full editing procedures could be explored in further research.

Under the Project Connect deterministic editing and imputation rules, per year missing values of gender are imputed using the school roster data for that year. Response patterns that then include both M and F answers were examined after the fourth year of data collection. The yearly imputation of gender could result in an inconsistent response pattern if

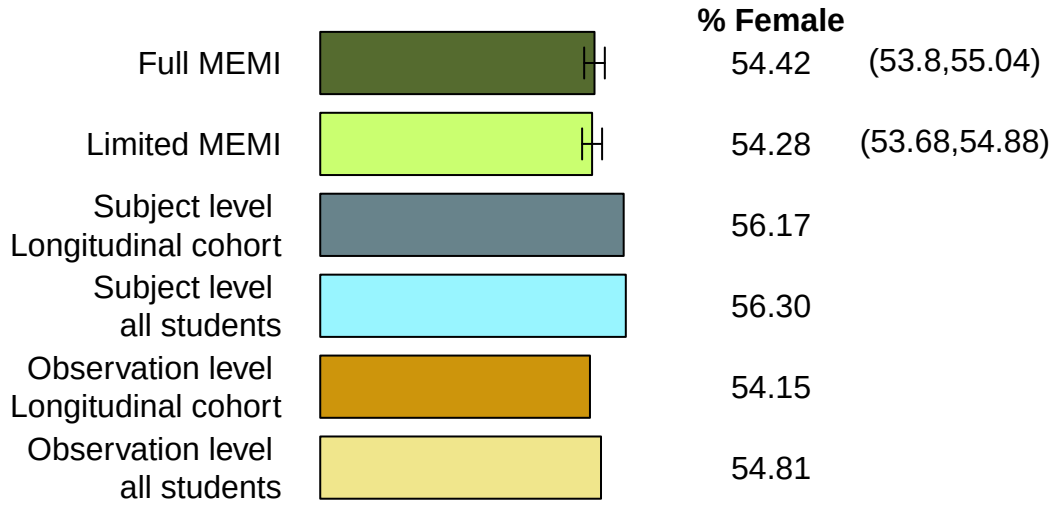


Figure 4.6: Comparison of several methods for calculating the percent female.

the roster gender data obtained from the school is incorrect. It is my experience that data obtained from the school is also subject to inconsistencies and should not be considered a gold standard to compare the self-report data to. The PC longitudinal edit rule for a binary IRM is a simple majority wins. If a student reported FMMM across years then that student is treated as male. If there is a tie (MMFF or MF) then school roster data is used as the tie breaker if available and also consistent. Because of these rules, the school roster data can have an undue influence on the final determination of a student's gender.

4.3 Example 2: Analysis of Birthplace

Next I analyze birthplace reported over time. Birthplace is used to determine the student's generation of immigration, which is used as a measure of acculturation. A student is considered to be first generation if they and their parents were born outside of the US, second generation if the student was born in the US but both parents were not, and third generation if both the student and at least one of their parents were born inside the US. The Project Connect survey asks students to specify their birthplace each year by selecting one of 8 options: United States, Mexico, El Salvador, Guatemala, China, The Philippines, Ko-

rea or Other. To model birthplace as a binary IRM, this variable has been dichotomized into "United States born" (1), and "Foreign Born" (0). This is not to say that inconsistent reports between other countries (such as China and Mexico) does not exist, but for the purposes of creating the generation of immigration variable the only distinction that matters is if the student is born inside versus outside the US.

4.3.1 Data

To model the probability λ_i a student was born in the US, I use a three level categorical ethnicity variable represented as three indicator variables; $\mathbf{w}_i = (\text{Hispanic/Latino}_i, \text{African-American}_i, \text{Other}_i)'$ without an intercept. This variable was defined as the value first reported by the student. To model the reporting probability π_{ij} that a student reports being born in the US given their true birthplace, I use FEMALE, an indicator of being female, and standardized AGE. I feel that gender could affect how a student answers the question of birthplace, it should not predict the actual birthplace. For purpose of illustrating methodology I use the Project Connect deterministically edited version of FEMALE, as defined in Section 3.5.1. This variable is missing on some students so they are excluded. These two variables plus an intercept make up the response level covariates $x_{ij} = (1, \text{AGE}_{ij}, \text{FEMALE}_{ij})'$. Ethnicity and age are fully observed predictors, records missing a reported birthplace were excluded to result in an analysis sample size of $N = 36,040$ surveys on $n = 26,438$ students. This illustrates an analysis using complete case data only. The model presented in Chapter 7 eliminates this problem by fitting the model at the current iteration using the complete (edited and imputed) data from the previous iteration.

4.3.2 Prior Distributions and Simulation Settings

Vague normal priors are placed on all regression coefficients. The prior means $\mathbf{m}_\alpha$ and $\mathbf{m}_\beta$ are set at $(5, 0, 0)'$ and $(5, 0, 0)'$ respectively. This reflects the prior belief that there is a high probability that a student will correctly report their birthplace, and that a priori gender and age have no expected effect on the reporting accuracy. The prior mean $\mathbf{m}_\gamma = (1.38, 2.2, 2.2)'$

is set to reflect the prior belief that Hispanic students have an 80% chance of being born in the US, while African-Americans and all others are set at 90%. Prior variances are constructed using equations (4.22a) and (4.22a) with $n_0 = n_1 = 5$,

Simulation settings. A total of $s = 5$ parallel MCMC chains are run with Phase 1 consisting of $D = 3$ blocks of $m = 1,000$ iterations discarded as burn-in. Phase 2 simulation is run for $M_2 = 40,000$ additional iterations per chain, retaining every $k = 40^{\text{th}}$ iteration. This results in a final sample size of $\tilde{n} = 5,000$. Convergence was monitored using trace, density and Gelman-Rubin diagnostic plots.

4.3.3 Results

Table 4.4 gives summary statistics including the means and standard deviations of the posterior samples of the regression coefficients and the average probability a student is born in the US. Odds Ratios (OR) or probabilities (p in *italics*) with corresponding 95% intervals are included.

The probability of being born in the US for a Hispanic student is .795 (95%PI .787, .802), for an African-American student .974 (.967, .980), and for students of other ethnicities .66 (.64, .68). The average probability that a domestically born student will correctly report their birthplace is .990 (.986, .993), and the average probability that a foreign born student will correctly report their birthplace is 1-.04 = .96 (.95, .97). Being female is associated with a larger odds of accurately reporting birthplace. US born females have 1.3 times the odds of reporting they were born in the US compared to males, and foreign born females have 1/.365 = 2.7 times the odds of reporting they were not born in the US compared to males. Parallel to what was seen with gender, the probability of accurately reporting birthplace decreases as age increases for those born in the US, and increases with age for those born elsewhere.

Missing and inconsistent values for birthplace are imputed or edited $M = 20$ times using Equation (4.20). Under the limited editing procedure the percent of students in the data

Parameter	Interpretation	Mean	SD	OR/p	2.5%	97.5%	$p(\theta_j > 0)$
$\bar{\lambda}$	p(US Born)	0.801	0.002	-	0.797	0.805	1.000
γ_1	Hispanic	1.354	0.023	*0.795*	*0.787*	*0.802*	1.000
γ_2	African-American	3.630	0.128	*0.974*	*0.967*	*0.980*	1.000
γ_3	Other Ethnicity	0.662	0.042	*0.660*	*0.642*	*0.678*	1.000
US Born							
α_1	Intercept	4.572	0.178	*0.990*	*0.986*	*0.993*	1.000
α_2	Female	0.276	0.161	1.318	0.960	1.819	0.959
α_3	Age (standardized)	−1.458	0.170	0.233	0.165	0.324	<0.001
Foreign Born							
β_1	Intercept	−3.175	0.178	*0.040*	*0.028*	*0.055*	<0.001
β_2	Female	−1.006	0.254	0.365	0.217	0.593	<0.001
β_3	Age (standardized)	−0.966	0.112	0.381	0.305	0.472	<0.001

Table 4.4: Summary of the posterior distribution for the IRM birthplace model parameters. Odds Ratios (OR) or probabilities (p in *italics*) with corresponding 95% posterior intervals are included.

set used for this example that are born in the US is 79.36% (78.72%, 79.99%), and 78.65% (78.15%, 79.15%) under the full edit.

4.3.4 Discussion

Sample Size and Missing Data. Table 4.5 looks at the bivariate combination of missing and inconsistent responses for birthplace and gender. Compared to the reports on gender, students were more likely to inconsistently report their birthplace over time; 165 (0.7%) students inconsistently reported their birthplace but not gender, 52 (0.9%) inconsistently reported their gender but not birthplace, and 5 (0.08%) inconsistently reported both. More people were excluded from the modeling stage due to missing gender (245) than missing birthplace (194). A total of 364 (1.4%) students had their birthplace edited or imputed.

The Project Connect data set has more variables that could have been used as predictors

		Gender			
		Missing	Consistent	Inconsistent	Total
Birthplace	Missing	26	168	0	194
	Consistent	219	26216	52	26487
	Inconsistent	0	165	5	170
	Total	245	26549	57	26851

Table 4.5: Frequency of missing and inconsistently reported gender and birthplace.

of birthplace. These include the language the survey was taken in, the language(s) spoken at home, and the mother's and father's birthplaces. However, not only is the percentage of missing data in these additional covariates high, the model of how the parental birthplaces relate to the students' birthplace is complicated. The prior belief that a student will not know his or her parent's birthplace is non-negligible. Neither is the possibility that the student might not know his or her parent's birthplace one year, yet would know it the next.

Figure 4.7 shows that the point estimates for the percent of students born in the US are fairly similar regardless of how the estimate is calculated. However, the MEMI estimates properly account for the error in the editing and imputation process and show error bars around this estimated percent.

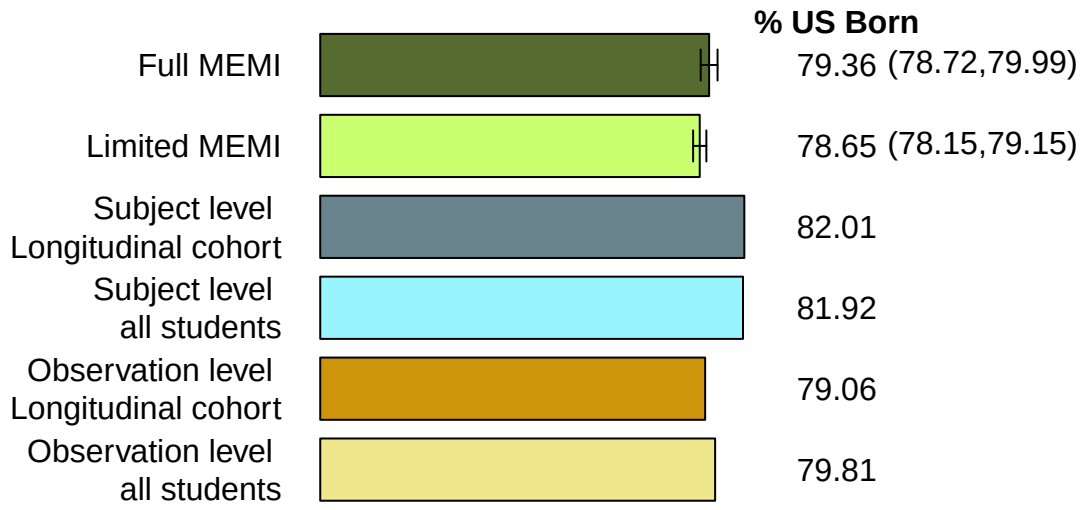


Figure 4.7: Comparing %US born using several different calculations methods.

4.4 Discussion

This chapter introduced the inconsistent repeated measures (IRM) model and provided two examples of fitting this model and the multiple imputation and editing of the inconsistently reported measures. If no subject level covariates are available, a modification to this model can include placing a Beta prior on λ, the probability of the underlying true value being 1.

This model does not consider the temporal nature of the sampling and considers each survey to be independent repeated measures on the same student. Disregarding the ordering of the responses is reasonable because this variable has an underlying time-fixed truth. There is only one underlying true value for each student. This model also only edits and imputes one variable, the student's birthplace and uses a version of gender that was subject to missing and deterministically edited data.

Chapter 7 uses the SyBRMICE procedure to demonstrate a method to edit and impute birthplace using a gender variable that has also been subject to data editing and imputation.

CHAPTER 5

Modeling Inconsistent Monotone Longitudinal Responses

Many longitudinal surveys ask about lifetime behaviors: "Have you ever..." questions. There are three consistent longitudinal response patterns for this type of variable; all No (0), all Yes (1), or with a single change from always No until the change and always Yes (0 to 1) after the change. These variables have a structured trend; they are allowed to change but only in a monotonic manner (0 to 1). An inconsistent response pattern is when the responses go from Yes to No (1 to 0), which include having multiple change points (No, Yes, No). This chapter develops a method for modeling, imputing and editing inconsistent monotonic longitudinal responses to a repeated binary question with a structured trend.

I use a survival analysis approach to model the time where the change point occurs, the *time of event*, t_i for student $i = 1, \ldots, n$ as a function of related predictors $\mathbf{x_i}$. Missing and inconsistent survey data y_{ij} are then imputed or edited using samples of posterior predicted values $t_i^{(\ell)}$.

5.1 Data Management

I construct a balanced data set where all students have equal numbers of surveys; the data set is filled out into a balanced data set with missing values if a student has missing data that year. Then all students have 7 surveys, one for each grade 6 through 12. For example, if a student only participated in 8^{th} and 9^{th} grades, rows for 6^{th} , 7^{th} , 10^{th} , 11^{th} and 12^{th} grades are added to the data. When initially expanded, the newly introduced rows in are completely missing on all variables except grade and need to be filled in. For clarity when using this

balanced structure I let $j = 6, \ldots, 12$ index grade level instead of survey number. Then let y_{ij} be the value of the lifetime question for student i at grade j. This balanced data structure will be used again in Chapter 7, so I describe which variables can be deterministically filled in and how this is done at this time, but I do not use all of the variables described until later. The resulting dataset has an additional 151,384 surveys, for a total of 187,957 records.

Ethnicity is a subject level variable, so the data from the observed records are copied to all other records for the same student. AGE and study wave are incremented up or down according to the location of the missing data relative to the observed data. Missing WEAP and FIGHT are imputed as never carrying a weapon in the past 30 days, and not fighting in the past 12 months (imputed values of 0). The presence of an SBHC is completely determined by the school ID, so missing data on school ID is imputed using last observation carried forward (LOCF) and first observation carried backward (FOCB). If no data during the high-school years was available then the school ID is imputed as the school ID that corresponds to the feeder middle school the student reported attending. The value of intervention depends on the study wave and school ID.

5.2 Modeling Time of Event

The time of event t_i is modeled using an interval censored regression model

$$P(t_{1i} > t_i > t_{2i}) = \Phi(t_{2i}|\mathbf{x}_i'\boldsymbol{\theta}, \sigma^2) - \Phi(t_{1i}|\mathbf{x}_i'\boldsymbol{\theta}, \sigma^2) \tag{5.1}$$

where Φ is the normal cumulative distribution function, and the pair (t_{1i}, t_{2i}) represent the lowest and highest possible values for the event time for student i. Proper priors are assigned to the regression parameters $\boldsymbol{\theta}$ and the variance σ^2

$$\begin{aligned} \boldsymbol{\theta} &\sim N(m, V), \\ \sigma^2 &\sim IG(a, b). \end{aligned} \tag{5.2}$$

The model is fit using the `MCMCglmm` (Hadfield, 2010c) package in `R` which uses an M-H sampling algorithm which adapts during a specified burn-in period.

The pair of lowest and highest possible values, (t_{1i}, t_{2i}), for the event time depend on

the type of censoring. A student with a 1 on the first survey response is left censored because the event occurred before the first time point. A student with all observed $y_{ij} = 0$ is right censored because the event did not occur prior to the last observed time point. If the change point occurs during the study period, it is considered interval censored because this survival model treats time as continuous but we only observe discrete survey time points. The censoring time points are then

$$(t_{1i}, t_{2i}) = \begin{cases} (-\infty, g_{ij}) \text{ for left censored,} \\ (g_{ij}, \infty) \text{ for right censored,} \\ (g_{ij^1}, g_{ij^2}) \text{ for observed events,} \end{cases} \tag{5.3}$$

where g_{ij} is the grade level at the event time for subject $i = 1, \ldots, n$, g_{ij^1} is the grade where the last 0 was observed, and g_{ij^2} is the grade where the first 1 was observed. So if the reports for $j = 9, 10, 11$ are (0,0,1) the interval is (10,11). If the 10^{th} grade report were missing, the interval would be (9, 11).

5.3 Monotone Editing and Missing Data Imputation

Let j^0 and j^1 be the grade of the first and last observed value for y_{ij} respectively. Then there are four distinct groups of missing data to be imputed: Prior to the first observed survey when the value on that survey is 0, prior to the first observed survey when the value on that survey is 1, after the last observed survey when the value is 0, and after the last observed survey when the value is 1. Two of these can be deterministically imputed, the other two require modeling. Starting values for the missing data in these two groups will be discussed later.

If a student reported No (0) on their first survey, every survey prior to that one can be imputed deterministically as 0; if $y_{ij^0} = 0$ then set $y_{ij} = 0$ for all $j < j^0$ (FOCB). If the last observed y_{ij} is a 1 then impute a 1 for all later surveys (LOCF). For example if a student participates in 9^{th} and 10^{th} grades only, with $y_{i9} = 0$ and $y_{i10} = 1$, then set $y_{ij} = 0$ for $j = 6, \ldots, 8$ and $y_{ij} = 1$ for $j = 11, 12$. This process deterministically imputed 56,588 values

for y_{ij}.

Missing data indicators $M_{ij} = 1$ are assigned if y_{ij} is not observed. Observations that are deterministically imputed are not considered missing. Erroneous data indicators are assigned to all 7 records on each student i reporting an inconsistent pattern. If a 0 was observed after a 1 was reported, alternatively if $y_{ij} = 1$ and $y_{ij'} = 0$ for any $j < j'$ then $E_i = 1$.

Consider the ℓ^{th} draw $\boldsymbol{\theta}^{(\ell)}$ from the posterior density $p(\boldsymbol{\theta}|t, \mathbf{x})$. The predicted time of first sex $t_i^{(\ell)}$, is drawn from a truncated normal distribution with mean $\mu_i^{(\ell)} = \mathbf{x_i}'\boldsymbol{\theta}^{(\ell)}$ and variance $\sigma^{2(\ell)}$

$$t_i^{(\ell)} = \Phi^{-1}\left[\Phi\big(\frac{a_i - \mu_i^{(\ell)}}{\sigma^{2(\ell)}}\big) + u_i\Big(\Phi\big(\frac{b_i - \mu_i^{(\ell)}}{\sigma^{2(\ell)}}\big) - \Phi\big(\frac{a_i - \mu_i^{(\ell)}}{\sigma^{2(\ell)}}\big)\Big)\right], \tag{5.4}$$

where Φ^{-1} is the inverse normal CDF, the truncation limits $(a_i, b_i) = (t_{1i}, t_{2i})$, and $u_i \sim U(0,1)$ is a uniform random variable. This predicted value is used to edit and impute the response level values for y_{ij} as follows. Then the ℓ^{th} imputed or edited value is

$$y_{ij}^{(\ell)} = \begin{cases} y_{ij} & \text{if } M_{ij} = 0 \text{ and } E_i = 0 \\ 0 & \text{if } (M_{ij} = 1 \text{ or } E_i = 1) \text{ and } j < t_i^{(\ell)} \\ 1 & \text{if } (M_{ij} = 1 \text{ or } E_i = 1) \text{ and } j \geq t_i^{(\ell)}. \end{cases} \tag{5.5}$$

If the observed value is not part of an inconsistent response pattern then the true value equals the observed value. For any missing values or inconsistent response patterns the true value is set at 0 for all grades lower than the predicted grade of first sex, and set at 1 for all grades higher than the predicted grade.

5.4 Example: Ever Had Sex

The analysis sample using the balanced data structure is 183,512 surveys on 26,216 students not missing the Project Connect deterministically edited versions of gender or birthplace. Of these, 231 (0.8%) students reported an inconsistent response pattern for ever having sexual intercourse (`I1`) on 1,617 surveys. The value for `I1` is missing on 96,767 (52.7%) observations. The total number of surveys then subject to editing or imputation due to inconsistent or missing data is 98503 (53.7%).

Ethnicity, age of entry into the study, gender and birthplace are the variables used to predict the event time, grade at first sex.

$$\mathbf{x_i} = (1, \text{AA}_i,\ \text{Other}_i,\ \text{Age at entry}_i,\ \text{FEMALE}_i,\ \text{US}_i)'$$

The prior mean vector $m = (12, 0, 0, 0, 0, 0)'$ for the regression coefficients $\boldsymbol{\theta}$ is set to reflect the belief that the average grade at first intercourse is 11th grade, and that all other variables a priori are not expected to be associated with grade at first sex. The prior covariance matrix is set as the identity matrix $V = I_6$. The hyper-parameters a and b for the prior distribution on σ^2 are $a = b = .001$.

5.4.1 Starting Values

Starting values $y_{ij}^{(0)}$ for the missing y_{ij} are created by using the marginal sample probabilities of grade at first sex calculated from the sample of observed y_{ij} and shown in equation (5.6). If the first observed response is a 1, $y_{ij^0} = 1$, earlier responses are initialized using these probabilities. If the last observed response is a 0, $y_{ij^2} = 0$, later responses are initialized in the same manner. The value of 13 represents anything later than 12^{th} grade. While t is a continuous measure the starting values are imputed using only the integer portion. Draw $u_i \sim$ Uniform(0,1). Then for all $j < j^0$ and $M_{ij} = 1$, set

$$t_{ij}^{(0)} = \begin{cases} 6 \text{ if } u_i \leq .01 \\ 7 \text{ if } .01 < u_i \leq .02 \\ 8 \text{ if } .02 < u_i \leq .04 \\ 9 \text{ if } .04 < u_i \leq .09 \\ 10 \text{ if } .09 < u_i \leq .16 \\ 11 \text{ if } .16 < u_i \leq .25 \\ 12 \text{ if } .25 < u_i \leq .38 \\ 13 \text{ if } u_i > .38 \end{cases}, \tag{5.6}$$

then equation (5.5) is used to impute the missing y_{ij} as 0 or 1. Starting values for the regression parameters $\boldsymbol{\theta}$ were set by the `MCMCglmm` program defaults.

Parameter	Mean	SD	2.5%	97.5%	$P(\theta > 0)$
Intercept	9.968	0.280	9.440	10.537	1.000
African-American	−0.869	0.074	−1.016	−0.723	<0.001
Other Ethnicity	0.691	0.078	0.538	0.844	1.000
Age at entry	0.024	0.017	−0.011	0.057	0.915
Female	0.904	0.048	0.812	1.001	1.000
US Born	−0.042	0.059	−0.156	0.077	0.242
Sigma (σ)	2.721	0.050	2.628	2.821	-

Table 5.1: Summary of the posterior distribution for the grade of first sex survival model parameters with 95% posterior intervals and p-values.

5.4.2 Modeling Results

`MCMCglmm` was run for 51,000 iterations retaining every 10^{th} iteration and discarding the first 1,000 as burn-in, resulting in a final sample size of 5,000. Convergence was monitored using trace and density plots. The model converged rapidly, within the first thousand iterations.

Table 5.1 gives summary statistics for the IML model of grade at first sex, including posterior means, standard deviations and 95% posterior intervals (PI) for the regression coefficients $\boldsymbol{\theta}$ and standard deviation σ. A foreign-born male Hispanic student is expected to have engaged in sexual intercourse in the 10^{th} grade, 9.97 (95%PI 9.44, 10.54). Being female postpones the grade of first sex by .904 (.812, 1.00) years, and the age when the student first participated in Project Connect was not associated with grade of first sexual intercourse.

Figure 5.1 shows trace and density plots for each of the regression coefficients. Prettier pictures could have been drawn manually, but I wanted to take this opportunity to demonstrate the default output from `plot.MCMCglmm`, which is quite sufficient for diagnostic purposes. The bottom two subplots with the label *units* represents the modeled variance parameter σ^2.

5.4.3 MEMI Results

One way to examine the percent of students who are sexually experienced without multiply counting those with repeated measures is to look at the percent of $y_{ij} = 1$ across grade levels. Figure 5.2 plots the percent calculated on the complete cases as red dots, and the percents combined across all 20 MEMI data sets as blue confidence intervals. This shows that the combined estimates tend to be a bit higher than the observed complete case percent, but the interval always covers the CC estimate.

When analyzing survey data on measures such as birth control use at last sex, or going to get a condom from someone on campus, the relevant population is the students who have ever had sex. For the Project Connect data, this means the subset of surveys where the response to "Have you ever had sex" is "Yes". For the multiply edited and imputed data sets, this is where $y_{ij} = 1$, and in the case of the complete case unedited data set, where $y_{ij} = 1$. These subsets still contain repeated measures from the same student, so subsequent analyses need to correctly account for the student clustering.

Figure 5.3 shows what the size of this subset would looks like if calculated on the three types of data structures discussed in this chapter. The proportion of surveys from students who have ever had sex is 34.12% of the complete cases, 30.85% (30.47%, 31.22%) of the surveys from the MEMI balanced data structures, and 34.76% (34.26%, 35.26%) from the MEMI data sets restricted to those surveys that were originally observed. The proportion calculated on the balanced structure is lower than the other two due to a large number of 0's deterministically imputed. The average then is calculated on more 6^{th} and 7^{th} graders than the other methods.

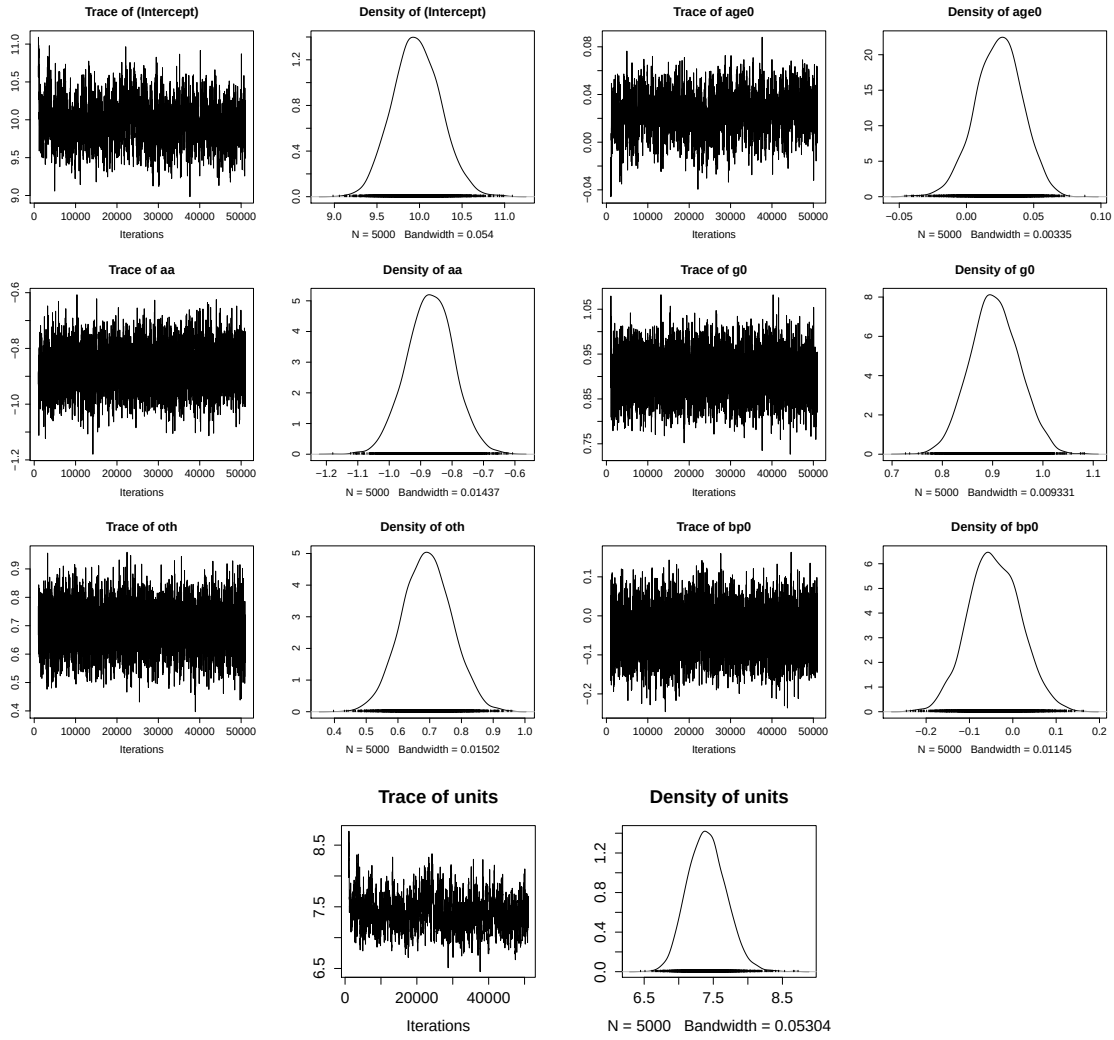


Figure 5.1: Posterior trace and density plots for the grade of first sex model regression coefficients and the variance parameter σ^2 (labeled as *units*) using the default plotting function from `MCMCglmm`.

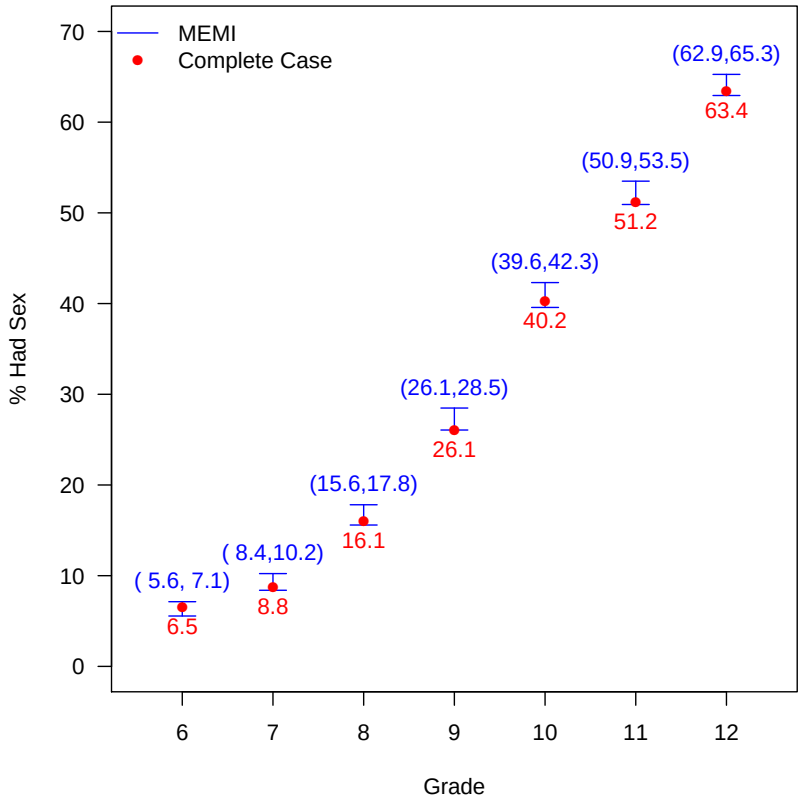


Figure 5.2: Percent of students who ever had sex by grade. Percents calculated on the complete case data set shown as red dots, confidence intervals for the percent combined across all 20 MEMI data sets shown in blue.

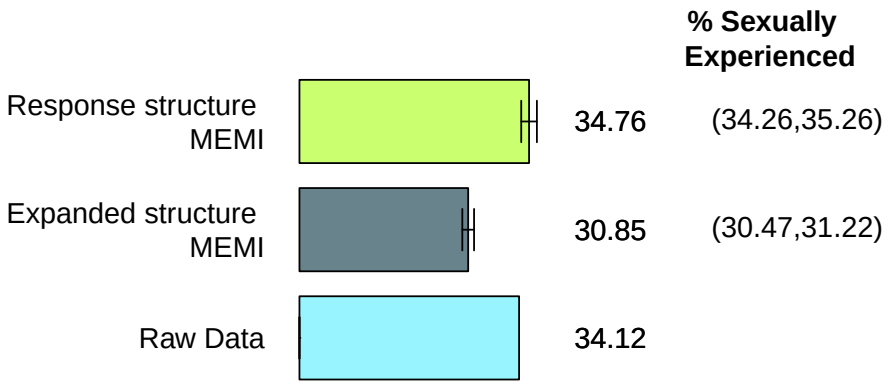


Figure 5.3: Comparison of results from three methods for calculating the percent of observations on sexually experienced students.

5.5 Discussion

This chapter introduced a method to stochastically edit and impute inconsistent and missing data for a variable that has a structured trend. Caution should be given to interpreting the regression coefficients with respect to interpolation into the months within the year of first sex. All students were surveyed each spring, but it ranged from February to June. As with the example of the foreign born Hispanic male on average having sex at grade 9.9, this is not suggesting that it occurred in the summer prior to the start of the official 10^{th} grade academic year. All that can be determined by the reports of sexual experience is if the event occurred prior to the survey date of that academic year.

The predicted time of first sex had a similar distribution to the observed grade at first sex. There exist other variables in the Project Connect dataset that would be appropriate to used as predictors, but given the amount of missing and potentially inconsistent data in those variables, I chose to keep this example simple to serve as an illustration of the model. A more thorough approach to modeling variables subject to missing and inconsistent data using variables also subject to missing and inconsistent data is demonstrated in the SyBRMICE procedure of Chapter 7. Additionally further work should include modifying the editing model to make minimal changes.

CHAPTER 6

Modeling Inconsistent Multivariate Responses

This chapter presents a method for modeling and editing inconsistent multivariate responses (IMV). This type of inconsistency occurs when responses between two or more questions give conflicting information. For example if the student reports never having sex in their lifetime, but then on a following question reports having sex in the past three months, these responses would be inconsistent.

If a two- or multi-way contingency table is created using variables with conflicting responses, any cell that results in an IMV should be a structural zero: a cell corresponding to an inconsistent response combination should have exactly zero records in it. Consider an example of two binary variables (Y_1, Y_2) both taking on values 0 and 1 where a response combination of $(Y_1, Y_2) = (1, 0)$ is a structural zero. Let $P(Y_1 = 1) = \phi_1$ and $P(Y_2 = 1|Y_1 = 0) = \phi_2$, then the joint probability distribution of Y_1 and Y_2 is shown in Table 6.1.

This chapter provides a procedure that defines a model to estimate the probabilities ϕ_1 and ϕ_2, and a model to stochastically edit the inconsistent responses between these variables. I apply this procedure to the knowledge and utilization variables of the Condom Availability Program in Project Connect high schools, and compare the results to several alternative editing methods.

		Y_2	
		0	1
Y_1	0	**A**: (1-ϕ_1)(1-ϕ_2)	**B**: $(1-\phi_1)\phi_2$
	1	**C**: 0	**D**: ϕ_1

Table 6.1: Joint distribution of two binary variables Y_1 and Y_2 with a structural zero at **C**: (1,0). The **A**, **B**, **C**, **D** are cell labels.

6.1 Model Specification

Let y_{i1} and y_{i2} be data observed on binary variables Y_1 and Y_2 respectively from $i = 1, \ldots, n$ students. Let y_{i1} be distributed as Bernoulli with probability ϕ_{i1} for all $i = 1, \ldots, n$ observations, and for the n_0 observations where $y_{i1} = 0$ let y_{i2} be distributed as Bernoulli probability ϕ_{i2}. For the remainder $n - n_0$ observations with $y_{i1} = 1$, y_{i2} is constrained to be 1. Let $\mathbf{x_1}$ be the $n \times p$ matrix of data on p covariates from n students used to predict y_1 and let $\mathbf{x_2}$ be the $n_0 \times q$ matrix of data on q covariates from n_0 students used to predict y_2 when $y_1 = 0$. Since Y_1 and Y_2 are so closely related, $\mathbf{x}_1$ and $\mathbf{x}_2$ likely have variables in common.

The IMV model is then written as

$$\begin{aligned} y_{i1} &\sim \text{Bernoulli}(\phi_{i1}) \\ \phi_{i1} &= \Phi(\mathbf{x_{i1}}'\boldsymbol{\theta}_1) \\ (y_{i2}|y_{i1} = 0) &\sim \text{Bernoulli}(\phi_{i2}) \\ \phi_{i2} &= \Phi(\mathbf{x_{i2}}'\boldsymbol{\theta}_2) \\ (y_{i2}|y_{i1} = 1) &= 1 \text{ with probability } 1, \end{aligned} \tag{6.1}$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are vectors of unknown regression parameters of length p and q respectively. This hierarchical model is fit using data augmentation and Gibbs sampling as described by Albert and Chib (1993). This method for generating samples from the joint posterior of ϕ_{i1} and ϕ_{i2} is described next.

6.1.1 Sampling Algorithm

The Albert-Chib method (Albert and Chib, 1993) introduces a vector of n continuous variables $Z = (Z_1, \ldots, Z_n)'$. The Z_i have independent normal distributions with mean $\mathbf{x_i}'\boldsymbol{\theta}$ and variance 1, where X is the $n \times p$ design matrix of predictor data with i^{th} row $\mathbf{x_i}$ and $\boldsymbol{\theta}$ the vector of corresponding unknown regression coefficients. The latent variables Z are connected to the variable Y in that $y_i = 1$ if $Z_i > 0$ and $y_i = 0$ if $Z_i \leq 0$. Then

$$\phi_i = P(y_i = 1) = P(Z_i > 0) = \Phi(\mathbf{x_i}'\boldsymbol{\theta}),$$

for $i = 1, \ldots, n$.

The key is to consider this situation as a missing data problem where Z_i is a latent variable and only observed through y_i. The conditional posterior distribution for the latent variable Z_i given $\boldsymbol{\theta}$ and $\mathbf{y_i}$ is

$$\begin{aligned} [Z_i|\boldsymbol{\theta}, y_i] &\sim \mathcal{N}(\mathbf{x_i}'\boldsymbol{\theta}, 1)I(Z_i > 0), \text{ if } y_i = 1, \\ [Z_i|\boldsymbol{\theta}, y_i] &\sim \mathcal{N}(\mathbf{x_i}'\boldsymbol{\theta}, 1)I(Z_i \leq 0), \text{ if } y_i = 0, \end{aligned} \tag{6.2}$$

where Z_i is simulated from a truncated normal distribution (5.4), with the bounds (a_i, b_i) on the truncation being either $(0, \infty)$ if $y_i = 1$, or $(-\infty, 0]$ if $y_i = 0$. If an informative multivariate normal prior, $\mathcal{N}_p(\boldsymbol{\theta}^0, V_0)$, is placed on the regression coefficients $\boldsymbol{\theta}$, the conditional posterior distribution for $\boldsymbol{\theta}$ is

$$[\boldsymbol{\theta}|Z, X] \sim \mathcal{N}_p(\boldsymbol{\theta}^1, V_1), \tag{6.3}$$

where

$$\boldsymbol{\theta}^1 = (X'X + V_0^{-1})^{-1}(X'Z + V_0^{-1}\boldsymbol{\theta}^0), \tag{6.4}$$

$$V_1 = (X'X + V_0^{-1})^{-1}. \tag{6.5}$$

Gibbs sampling is then used to generate samples from the joint posterior density of Z and $\boldsymbol{\theta}$ by alternating draws from the conditional posterior distributions of the Z_i's and the $\boldsymbol{\theta}$'s.

After introducing a second set of latent variables $Z_i^*, i = 1, \ldots, n_0$ that are associated with y_{i2} (when $y_{i1} = 0$) in the same manner as described for Z_i and y_i, model (6.1) can be

written as

$$\begin{aligned} y_{i1} &\sim \text{Bernoulli}(\phi_{i1}) \\ \phi_{i1} &= P(y_{i1} = 1) \\ &= P(Z_i > 0) \\ &= \Phi(\mathbf{x_{i1}}'\boldsymbol{\theta}_1) \end{aligned} \tag{6.6a}$$

$$\begin{aligned} (y_{i2}|y_{i1} = 0) &\sim \text{Bernoulli}(\phi_{i2}) \\ \phi_{i2} &= P(y_{i2} = 1|y_{i1} = 0) \\ &= P(Z_i^* > 0|y_{i1} = 0) \\ &= \Phi(\mathbf{x_{i2}}'\boldsymbol{\theta}_2) \end{aligned} \tag{6.6b}$$

$$(y_{i2}|y_{i1} = 1) = 1 \text{ with probability } 1 \ .$$

The regression model (6.6a) is fit first and a posterior sample of $\boldsymbol{\theta}_1$ is drawn. Model (6.6b) is then fit on the subset of data where $y_{i1} = 0$ and a posterior sample of $\boldsymbol{\theta}_2$ is drawn. Next I discuss how these model results are used to generate stochastic edits for the inconsistent multivariate responses between Y_1 and Y_2.

6.2 Multiple Editing Procedures

Similar to the inconsistent repeated measures editing methods introduced in Chapter 4, the probabilistic editing method to correct inconsistent multivariate responses has a *limited* version where only those who are observed to have made an error are edited, and a *full* version that is more general and gives everyone the chance to have made a reporting error.

Recall the 2 x 2 cross-tabulation of Y_1 and Y_2 depicted in Table 6.1. The editing procedures used to correct this IMV re-allocate records out of the inconsistent cell **C** and back into one of the valid (**A**, **B**, or **D**) cells using normalized multinomial probabilities that differ between the limited and full editing procedure. These editing procedures also consider the probability of providing IMV reports between Y_1 and Y_2. Let π_1 be the probability of incorrectly reporting Y_1 and π_2 be the probability of incorrectly reporting Y_2. In this example these two probabilities are estimated from the data by dividing the number of inconsistent

		Y_2	
		No	Yes
Y_1	No	**A**: $(1-\phi_{i1})(1-\phi_{i2})\pi_1$	**B**: $(1-\phi_{i1})\phi_{i2}\pi_1\pi_2$
	Yes	**C**	**D**: $\phi_{i1}\pi_2$

Table 6.2: Unnormalized multinomial re-allocation probability distribution under the limited editing procedure to correct IMVs.

reports by the number of responses for both variables

$$\pi = \pi_1 = \pi_2 = \frac{\#\text{ in Cell } \mathbf{C}}{2N} = 0.018, \tag{6.7}$$

then treating π as known.

Limited Edit. The limited editing procedure only changes records that are in cell **C**. If the true values for (Y_1, Y_2) were **A**:(No, No) then observing a report combination of **C**:(Yes, No) was a result of the student making a single reporting error in Y_1. The probability of this occurring is calculated as the probability of being in cell **A** $p_{Ai} = (1-\phi_{i1})(1-\phi_{i2})$ times the probability of mis-reporting Y_1, π_1. Similarly if the true combination was **D**:(Yes, Yes), then the probability of a record observed in cell **C** is $p_{Di} = \phi_{i1}\pi_2$. However if the true combination was **B**:(No, Yes), then the inconsistent response combination **C** was a result of mis-reporting both Y_1 and Y_2. The probability of this occurring is $p_{Bi} = (1-\phi_{i1})\phi_{i2}\pi_1\pi_2$. These probabilities are displayed in Table 6.2 and after normalization, they serve as the multinomial re-allocation probabilities for an observation in cell **C**.

Full Edit. The full editing procedure reflects a model wherein each student has an underlying true cell that they belong to, which is potentially different from their observed cell. This means any combination of (Y_1, Y_2) could be a result of a reporting error and thus can be re-allocated back to any other valid cell, e.g. **A**$\rightarrow$ **B**, or **D**$\rightarrow$ **A**. I structure the subscripts as $p_{(OBS)(TRUE)i}$, where the first subscript denotes the cell the record was observed in, the second subscript denotes the cell that reflects the true combination of responses.

Observed Cell	True Cell **A**	**B**	**D**
A	$(1-\phi_{i1})(1-\phi_{i2})(1-\pi_1)(1-\pi_2)$	$(1-\phi_{i1})\phi_{i2}(1-\pi_1)\pi_2$	$\phi_{i1}\pi_1\pi_2$
B	$(1-\phi_{i1})(1-\phi_{i2})(1-\pi_1)\pi_2$	$(1-\phi_{i1})\phi_{i2}(1-\pi_1)(1-\pi_2)$	$\phi_{i1}\pi_1(1-\pi_2)$
C	$(1-\phi_{i1})(1-\phi_{i2})\pi_1(1-\pi_2)$	$(1-\phi_{i1})\phi_{i2}\pi_1\pi_2$	$\phi_{i1}(1-\pi_1)\pi_2$
D	$(1-\phi_{i1})(1-\phi_{i2})\pi_1\pi_2$	$(1-\phi_{i1})\phi_{i2}\pi_1(1-\pi_2)$	$\phi_{i1}(1-\pi_1)(1-\pi_2)$

Table 6.3: Multinomial re-allocation probability distribution under the full editing procedure to correct IMVs.

Table 6.3 lists all possible probabilities that an observation is found in one of three true cells along the top but is observed in one of the four cells on the left. Then the probability that an observation found in any cell on the left should be reallocated to the true cell along the top is proportional to the row probabilities given in the row where the observation is observed. To illustrate the full edit procedure, consider a record with a response combination of **A**:(No, No). The probability p_{AAi} in the table is the product of cell **A** being the true cell probability $(1-\phi_{i1})(1-\phi_{i2})$, times the probability that they did not make an error in either variable, $(1-\pi_1)(1-\pi_2)$. This is the first entry in the **A** row of table 6.3. Likewise p_{ABi} is calculated as the probability of truly being in **B**, $(1-\phi_{i1})\phi_{i2}$, times the probability of an error in Y_2 but not Y_1, $(1-\pi_1)\pi_2$. Lastly, p_{ADi} is the probability ϕ_{i1} of being in cell **D**, times the probability, $\pi_1\pi_2$, that an error was made in both variables.

Creating multiple edited data sets. Consider draws $\phi_{i1}^{(m)}$ and $\phi_{i2}^{(m)}$ from the posterior densities $p(\phi_{i1}|\mathbf{y}_{i1}, \mathbf{x_{i1}})$ and $p(\phi_{i2}|\mathbf{y}_{i2}, \mathbf{x_2})$ respectively for $m = 1, \ldots, M$. The reallocation probabilities are calculated for each student $i = 1, \ldots, n$ using either Table 6.2 normalized or the appropriate row of Table 6.3 normalized, and a vector $R_i^{(m)}$ from a multinomial distribution is drawn. For example under the limited editing procedure, $R_i^{(m)}$ is drawn as

$$R_i^{(m)} \sim \text{ Multinomial}(1, p_{Ai}^{(m)}, p_{Bi}^{(m)}, p_{Di}^{(m)}). \tag{6.8a}$$

The pair of edited values (y_{i1}^E, y_{i2}^E) is then

$$(y_{i1}^E, y_{i2}^E)^{(m)} = \begin{cases} (0,0) & \text{if } R_i^{(m)} = (1,0,0) \\ (0,1) & \text{if } R_i^{(m)} = (0,1,0) \\ (1,1) & \text{if } R_i^{(m)} = (0,0,1). \end{cases} \tag{6.8b}$$

Next is an example of applying this model to multiply edit potential inconsistent multivariate responses on the knowledge of and utilization of the Condom Availability Programs at Project Connect schools.

6.3 Condom Availability Program

Los Angeles Unified School District mandates Condom Availability Programs (CAPs) be in place in all of its high schools. One of the aims of Project Connect was to insure that CAPs were up and running in accordance with policy. One of the measures to assess the effectiveness of the CAP intervention was to ask students to report if they knew of someone on campus that gave out condoms. To measure utilization of the program the students were asked if they had ever gotten a condom from this person. The full text of these questions can be found in Appendix table A.4 where `G1` is the knowledge question and `G3` the question on utilization. These questions are worded so that the student does not have to know about the program by name, just whether or not they can get condoms from a person on campus.

The amount of condoms ordered each year by the school can be used as an additional measure of how well the CAP is functioning. DeRosa et al. (2012) showed that the number of condoms ordered by the high schools participating in Project Connect varied quite dramatically across study years. This is in part due to the presence (or absence) of a strong program champion, and because some schools took longer to achieve a successful and compliant CAP. Due to the volatile nature of the CAP, in combination with students switching schools, responses provided by students will be treated as independent across years, and a monotone structure is not imposed.

		Knowledge (Y_2)		
		No	Yes	Total
Utilization (Y_1)	No	**A:** 10,802 (55.3%)	**B:** 5,395 (27.6%)	16,197
	Yes	**C:** 673 (3.5%)	**D:** 2652 (13.6%)	3,325
	Total	11,475	8,047	19,522

Table 6.4: Distribution of response combinations between reported knowledge and utilization of the Condom Availability Program.

6.3.1 Data

Let Y_1 = utilization and Y_2 = knowledge. Table 6.4 shows the observed joint distribution of (Y_1, Y_2). More than half of students (55.3%) said they neither knew about nor utilized the CAP, nearly three in ten students (27.6%) said they knew of someone who gave out condoms but have not received one, and more than one in ten (13.6%) said they knew about, and have used the CAP. Almost 4% (673) of the 19,522 students gave an inconsistent response combination of not knowing that someone on campus gave out condoms, but that they have received condoms from this person. The n=19,522-673 = 18,849 consistent records are used to model ϕ_1 and $n_0 = 16,197$ records with $y_{i1} = 0$ are used to model ϕ_2.

Missing data is not considered at this time and only records with complete outcome and predictor data are used. A more thorough treatment of missing covariate data is developed in Chapter 7. The predictors used to model both Y_1 and Y_2 are: AGE (mean centered and standardized), presence of a School-Based Health Center (SBHC) on campus, an indicator of being in the intervention condition (INTERV), an indicator of ever having sexual intercourse (SEXP), gender (FEMALE) and being US born (US). As with the previous examples, the versions of gender and birthplace are the Project Connect deterministically edited versions. Then

$$X_1 = (1,\ \text{AGE},\ \text{SBHC},\ \text{INTERV},\ \text{SEXP},\ \text{FEMALE},\ \text{US})'$$

and X_2 has the same set of covariates, but excludes rows where $y_{i1} = 1$.

6.3.2 Priors and Simulation Settings

Multivariate normal priors were placed on the vectors of regression coefficients for utilization $\boldsymbol{\theta}_1$ and knowledge given no utilization $\boldsymbol{\theta}_2$,

$$\boldsymbol{\theta}_1 \sim \mathcal{N}\big((-2, 0, 1, .2, 1, 0, 0)', \frac{n}{25}(X_1'X_1)^{-1}\big) \tag{6.9a}$$

$$\boldsymbol{\theta}_2 \sim \mathcal{N}\big((-1, 0, 1, .2, .5, 0, 0)', \frac{n_0}{25}(X_2'X_2)^{-1}\big). \tag{6.9b}$$

The prior means on the intercepts were set based on a combination of results from trial runs and the prior belief that there is a low base probability of utilization or knowledge of the CAP.

Holding other variables at 0, there is a prior belief that students with an SBHC on campus have a $\Phi(-2+1)-\Phi(-2) = .14$ higher probability of utilizing, and a $\Phi(-1+1)-\Phi(-1) = .34$ higher probability of knowing about the CAP than those without. Students in intervention schools have around a .05 higher probability of knowing about and a .01 higher probability of utilizing the CAP compared to those in control schools. Lastly sexually experienced students are expected to have around a .15 higher probability of knowing about and getting a condom from the CAP compared to students who have not yet had sexual intercourse. Gender and birthplace are a priori not expected to be associated with either knowledge of or utilization of the CAP. If all covariates are set to 0 and the standardized age is allowed to vary, the min and max prior probabilities a student will know about the CAP is (.03, .68), and will go to get a condom from the CAP is (.002, .30).

Both regression models in Model (6.6) were fit using 11,000 iterations on each of $s = 5$ chains, discarding the first 1,000 and keeping every 10^{th} sample for a final posterior sample size of 5,000. Convergence was assessed using the diagnostic techniques described in Section 2.3.

6.3.3 Modeling Results

Define probabilities ϕ_1 and ϕ_2 to be the average probability of a student having gotten a condom from the CAP and knowing about the CAP

$$\begin{aligned} \phi_1 &= \frac{1}{n}\sum_{i=1}^{n}\phi_{i1}, \\ \phi_2 &= \frac{1}{n_0}\sum_{i=1|y_{i1}^{(\ell)}=0}^{n_0}\phi_{i2}. \end{aligned} \tag{6.10}$$

Table 6.5 gives summary statistics that describe the posterior distributions for the unknown parameters, including the average probabilities calculated in (6.10). Individual parameter estimates θ from a probit model do not have simple interpretation as they do in a logit model. Instead of Odds Ratios, I present the change in probability, Δp. This has the usual interpretation of the difference in the probability of the outcome y from 1 unit change in the covariate x, holding all other covariates at 0. This is calculated the same as when discussing the prior means. Since Φ is a nonlinear function, this calculation is done on each posterior sample with the mean, 2.5%tile and 97.5%tile shown in table 6.5.

Being male, older, attending an intervention school, or a school with an SBHC on campus, being sexually experienced and being born in the US are all associated with a greater probability of knowing about, or utilizing the CAP.

6.3.4 Editing Results

The data was edited $M = 20$ times under both the limited and full editing procedures. The results were combined back into the original data set to create Multiply Edited (ME) data sets which are indexed by $m = 1, \ldots, 20$. There are 20 ME's under the limited edit and 20 ME's under the full editing procedure.

Table 6.6 gives the reallocation distribution for the first 10 ME's under the limited editing procedure, which only edited the 673 inconsistent records in cell **C**. For example consider $m = 1$, ME#1; 485 (72.1%) were moved from $(Y_1, Y_2) = (1, 0)$ to (0,0), 183 (27.2%) were reallocated to (1,1), and the remaining 5 (.7%) were reallocated to cell **B**, (0,1). Table 6.7

compares the overall cell percents for the first 5 edits between the limited and full editing procedures. The results are similar between the two editing procedures, with the proportion in cell **A** ranging from 58.0% to 58.4% under the full editing procedure while under the limited edit, cell **A** has a range of 57.7% to 58.0%.

When looking at the cell percentages it is not clear how the results of the limited and the full editing procedures differ in practice. I take a step inward and look at how the raw data was changed on a univariate level. Table 6.8 shows how many, and in what direction, records were edited under each editing procedure for ME #1. The raw observed data for y_{i1} and y_{i2} are labeled $\texttt{G3}_i$, and $\texttt{G1}_i$ and the edited values are y_{i1}^E and y_{i2}^E respectively.

Table 6.8 shows that the limited editing procedure only changed records with G3=1 and G1=0. The top left corner table shows that 490 (72.8%) IMV records had their utilization variable changed from Yes to No, and the top right corner table shows that 188 (27.9%) IMV records had their knowledge variable changed from No to Yes. The bottom two tables show the results of the full editing procedure on a univariate scale. These tables show that 580 students had their response to the utilization question changed from Yes to No, and 44 from No to Yes. In addition, 160 students had their response to the knowledge question changed from Yes to No, and 262 from No to Yes. Univariate tables imply that the editing procedures favor editing people to Yes knowing but No utilization. However the joint editing tables clarify that a low percent of records are actually edited into the cell representing knowledge but no use.

Parameter	Mean	SD	Δp	2.5%	97.5%	$P(\theta > 0)$
Utilization(Y_1)						
ϕ_1	0.141	0.002	-	0.137	0.146	-
Intercept	−2.246	0.046	-	-	-	<0.001
Age (standardized)	0.078	0.022	0.003	0.001	0.005	1.000
SBHC	0.949	0.027	0.085	0.076	0.095	1.000
Intervention	0.166	0.025	0.006	0.004	0.009	1.000
Sexually Experienced	0.749	0.027	0.055	0.048	0.062	1.000
Female	−0.162	0.025	−0.004	−0.005	−0.003	<0.001
US Born	0.140	0.032	0.005	0.003	0.008	1.000
Knowledge | *No Utilization*$(Y_2 \| Y_1 = 0)$						
ϕ_2	0.333	0.004	-	0.326	0.340	-
Intercept	−1.139	0.034	-	-	-	<0.001
Age (standardized)	0.064	0.018	0.014	0.006	0.022	1.000
SBHC	0.810	0.021	0.244	0.228	0.259	1.000
Intervention	0.287	0.021	0.070	0.059	0.081	1.000
Sexually Experienced	0.101	0.023	0.022	0.0112	0.033	1.000
Female	−0.092	0.021	−0.018	−0.026	−0.010	<0.001
US Born	0.174	0.026	0.040	0.027	0.054	1.000

Table 6.5: Summary of the posterior distribution for the modeled parameters for the IMV CAP example. The one unit change in probability holding all other covariates at 0, Δp, and corresponding 95% posterior intervals for the change are included. The value of $p(\theta > 0|Y)$ is calculated by counting the number of times $\theta > 0$ and dividing it by the sample size.

	(Knowledge, Utilization)						
ME #	**A**(0,0)	**B**(0,1)	**D**(1,1)	ME #	**A**(0,0)	**B**(0,1)	**D**(1,1)
1	485	5	183	6	490	4	179
2	483	4	186	7	491	5	177
3	476	5	192	8	486	5	182
4	491	4	187	9	484	9	180
5	464	5	204	10	502	3	168

Table 6.6: Cell frequencies of (Utilization, Knowledge) for the reallocated observations after applying the limited multiple editing model (6.8). 10 out of the 20 MEMI's created are shown.

	Limited Edit Model			Full Edit Model		
ME #	**A**	**B**	**D**	**A**	**B**	**D**
1	57.82%	27.66%	14.52%	58.26%	27.46%	14.29%
2	57.81%	27.66%	14.54%	58.15%	27.50%	14.34%
3	57.77%	27.66%	14.57%	58.25%	27.50%	14.26%
4	57.85%	27.66%	14.50%	58.26%	27.51%	14.22%
5	57.71%	27.66%	14.63%	58.34%	27.45%	14.20%

Table 6.7: Valid cell percents for the first 5 edits after applying the limited and full editing IMV models.

		G3 (Utilization)			G1 (Knowledge)		
		y_{i1}^E				y_{i2}^E	
Limited Edit		No(0)	Yes(1)	**Limited Edit**		No(0)	Yes(1)
$\texttt{G3}_i$	No(0)	16197		$\texttt{G1}_i$	No(0)	11287	188
	Yes(1)	490	2835		Yes(1)		8047
		y_{i1}^E				y_{i2}^E	
Full Edit		No(0)	Yes(1)	**Full Edit**		No(0)	Yes(1)
$\texttt{G3}_i$	No(0)	16153	44	$\texttt{G1}_i$	No(0)	11213	262
	Yes(1)	580	2745		Yes(1)	160	7887

Table 6.8: Univariate IMV reallocation frequencies from ME #1 under the full and limited editing procedures. Entries in the diagonal cells in any sub-table are counts of unchanged values.

6.4 Comparison to Alternative Editing Methods

If no additional information from related covariates is available, valid cell counts (A, B, D) can be modeled as a multinomial outcome. Knowledge and utilization can be modeled jointly using

$$Y = (A, B, D) \sim \text{Multinomial}(\delta_A, \delta_B, \delta_D). \tag{6.11}$$

Assigning a uniform Dirichlet(1,1,1) prior on $\delta = (\delta_A, \delta_B, \delta_D)$ results in a conjugate posterior distribution Assigning a uniform Dirichlet(1,1,1) prior on $\delta = (\delta_A, \delta_B, \delta_D)$ results in a conjugate posterior distribution

$$g(\delta|Y) \sim \text{Dirichlet}(A + 1, B + 1, D + 1), \tag{6.12}$$

which can be sampled from directly. I created a posterior sample of size 5,000, and $M = 20$ draws from $g(\delta|Y)$were used to create a vector of multinomial probabilities, which were then used to re-allocate the inconsistent records back to one of the three valid cells.

I compare the proportion of students in each of the three valid cells after multiple different editing rules have been applied: complete case, where all inconsistent responses are discarded; deterministic, using the set of Project Connect editing rules; a joint probability model, and the conditional regression using the limited and full editing procedures. The complete case analysis drops all records with inconsistent responses. The deterministic editing rules used on the PC data are described in section 3.5.1, the joint probability model was just introduced and the conditional regression was introduced in Section 6.2.

Each row in Table 6.9 represents the cell proportions under each editing procedure. I produced $M = 20$ edited data sets under each of the multiple editing procedure described. Mean cell proportions and 95% intervals are calculated using Rubin's Rules (3.1) to combine cell counts across edited data sets. Proportions are calculated by taking the combined mean and 95% intervals for the counts and dividing by the sample size. Deterministic editing resulted in more people being classified as knowing about, but not using the CAP, whereas the regression model with either editing procedure resulted in more people being classified as both knowing about and getting a condom from the CAP than the other methods.

Editing Method	(Utilization, Knowledge) **A**(No, No)	**B**(Yes, No)	**D**(Yes, Yes)
Complete case	57.5	28.5	14.0
Deterministic	58.5	27.8	13.7
Joint	57.6	28.4	14.0
	(56.9-57.9)	(27.8-28.8)	(13.5-14.2)
Regression - Limited Edit	57.8	27.7	14.5
	(57.1-58.2)	(27.0-28.0)	(14.0-14.8)
Regression - Full Edit	58.2	27.5	14.3
	(57.5-58.6)	(26.8-27.8)	(13.8-14.5)

Table 6.9: Mean cell proportions and 95% intervals for the combination of Utilization and Knowledge under different editing models. All multiple editing procedures used $M = 20$ ME's.

This conditional regression model uses more subject-level information than the other editing models.

6.5 Discussion

The data used for Y_1 and Y_2 in this example are not the true raw versions as reported by the students on paper. The Project Connect analysis team noticed a discrepancy during a 2011 analysis that led them to (re)discover a change in the survey that caused previously unnoticed problems. In 2005, there was a skip pattern in place for the CAP section. If the student said "No" or "I don't know" to the question "Does someone at your school give out condoms?" they were instructed to skip to the next section. Not all students followed this skip pattern correctly. The text for this skip pattern was removed in 2006. Different skip patterns were observed between the baseline year compared to later years, so the deterministic imputation rules in place were re-examined and updated by the measurement team in 2011. The values

for G3 and G1 used in this dissertation are the updated versions of these variables. This also serves as an example of how edit and imputation rules are not necessarily 100% correct 100% of the time. The original rules that were put in place by the measurement team in 2006 (which led to the change in skip pattern wording after 2005) resulted in a lower number of students providing data on both G1 and G3, and a larger number of inconsistent combination of reports. This is not to say that the original recode rules were inferior in some way to the revised recode rules, just that different recode rules generate different results, which could result in different analysis conclusions.

CHAPTER 7

SyBRMICE

This chapter combines the ideas and models introduced earlier into a unified, novel method to handle multivariate missing and inconsistent data. Analyses performed using the multiple resulting data sets properly account for the additional variability due to uncertainty in the imputation and editing procedures. This method is called Sequential Bayesian Regression for Multiple Imputation and Conditional Editing (SyBRMICE). SyBRMICE adds stochastic editing to the Sequential Regression Multiple Imputation (SRMI) (Raghunathan et al., 2001) framework of section 3.2.1.

The next section outlines SyBRMICE, a sequential procedure where the results of one model feed into the next. I then define the indicator variables used to partition the data into the observed, missing and inconsistent portions of the data.

Two examples illustrate SyBRMICE. The first example in section 7.2 re-examines the inconsistent multivariate response combination of knowledge and utilization of CAP from Chapter 6. The second example in section 7.3 expands the first example to additionally incorporate Chapter 4's inconsistent repeated measures models for gender and birthplace, and Chapter 5's inconsistent monotone longitudinal response model of lifetime sexual experience. Section 7.4 examines how analysis results differ when models are based on the complete case data vs. the deterministically edited data vs. the multiple data sets created by the SyBRMICE process.

The two examples use variables that are only subject to a single edit. SyBRMICE can accommodate the more complex case where a single variable is subject to multiple editing models. However this requires a thoughtful setup regarding the order in which the imputation and editing models are applied.

7.1 SyBRMICE Notation and Algorithm

The SyBRMICE algorithm is sequential. Consider a set of variables $Y_1, \ldots, Y_P$ that are subject to missing and/or erroneous data. For each Y_p the SyBRMICE algorithm follows the general frameworks seen in earlier chapters. At step p of iteration ℓ a regression model $f_p(y_p^{(\ell-1)}|y_{-p}^{(\ell)}, \mathbf{X}_p, \boldsymbol{\theta}_p)$ is fit on the current values of Y_p, $y_p^{(\ell-1)}$ using fully observed predictors $\mathbf{X}_p$ specific to this step and $y_{-p}^{(\ell)}$, the current value of all other $Y_j, j \neq p$

$$y_{-p}^{(\ell)} = (y_1^{(\ell)}, \ldots, y_{p-1}^{(\ell)}, y_{p+1}^{(\ell-1)}, \ldots, y_P^{(\ell-1)}). \tag{7.1}$$

Posterior regression parameter estimates $\boldsymbol{\theta}_p^{(\ell)}$ are drawn and used in an imputation $g_p(y_p^I|y_{-p}^{(\ell)}, \mathbf{X}_p, \boldsymbol{\theta}_p^{(\ell)})$ model. If the vector of editing parameters $\boldsymbol{\phi}_p$ contain parameters not included in $\boldsymbol{\theta}_p$, those additional parameters are estimated and used in an editing $h_p(y_p^E|y_{-p}^{(\ell)}, \mathbf{X}_p, \boldsymbol{\phi}_p^{(\ell)})$ model. The results from both editing and imputation models are combined with the observed data to create a complete and consistent vector of imputed true responses $y_p^{(\ell)}$. The updated vector $y_p^{(\ell)}$ is then used as a predictor in the regression model for the other $Y_j, j \neq p$.

For example a regression model depending on unknown parameters $\boldsymbol{\theta}_1$ is fit to $y_1^{(\ell-1)}$ using predictors $(x_1, y_{-1}^{(\ell)})$, $\theta_1^{(\ell)}$ is drawn from its posterior distribution, and a complete true vector $y_1^{(\ell)}$ of true values is created. Then a model depending on $\boldsymbol{\theta}_2$ is fit to $y_2^{(\ell-1)}$ using predictors $(x_2, y_{-2}^{(\ell)})$. Regression parameters $\theta_2^{(\ell)}$ are drawn from the posterior and a complete true vector $y_2^{(\ell)}$ is created, which is then passed onto the next step. These steps continue until all P variables subject to missing and/or inconsistent data have been modeled, and the complete true data vector $y_P^{(\ell)}$ is passed back to the first regression model on $y_1^{(\ell)}$ to start the next iteration of the SyBRMICE algorithm.

SRMI has this same cyclical framework, with one main difference aside from the addition of the stochastic editing step. SRMI fits the regression models only on the observed data y_1 at each step, whereas SyBRMICE fits the regression model on the imputed and edited values from the previous iteration $y_p^{(\ell-1)}, p = 1, \ldots, P$. This process is outlined in detail in Algorithm 7.1.

Missing and Inconsistent Data Indicators To prepare the data for modeling I first create vectors of indicators of erroneous and missing data E_k and M_p as described in section 3.2. Recall that M_p is associated directly with Y_p, but E_k can involve more than 1 variable. Still, the missing and erroneous indicators are mutually exclusive. If the value of a variable is missing, it cannot be inconsistent.

For the simple case where a variable Y_p is subject to a single inconsistency E_p, the indicators can index the data such that each variable Y_p can be partitioned into n_p^{ok} observed consistent rows y_p^{ok}, n_p^E observed inconsistent rows y_p^E, and n_p^I unobserved, or missing, rows Y_p^I. So without loss of generalizability, $y_p = (y_p^{ok}, y_p^E, y_p^I)$ is the N-vector of data y_p where $n_p^{ok} + n_p^E + n_p^I = N$ for all $p = 1, \ldots, P$.

Next starting values for all unknown parameters are created. This includes regression parameters $\boldsymbol{\theta}_p^{(0)}$, and editing and imputation parameters $\boldsymbol{\phi}_p^{(0)}$ and $\boldsymbol{\theta}_p$. Also starting values for the missing $y_p^{I(0)}$ and inconsistent $y_p^{E(0)}$ data are generated, creating the starting data vector $Z_p^{(0)} = (y_p^{ok}, y_p^{E(0)}, y_p^{I(0)})$ for all p. Iteration $\ell = 1$ for variable Y_p of SyBRMICE is described in detail in Algorithm 7.1.

Not all editing and imputation procedures occur as separate steps. For example the inconsistent repeated measures (IRM) and inconsistent monotone longitudinal (IML) models impute missing and edit inconsistent data simultaneously, where the inconsistent multivariate response (IMV) model in this chapter imputes missing data then edits the inconsistent combinations. In the cases where the editing and imputation models are combined, steps 3 and 5 in algorithm 7.1 are combined into a single step that occurs after step 4.

Algorithm 7.1 Iteration ℓ of the SyBRMICE procedure.

For variable $Y_p, p = 1, \ldots, P$ at iteration ℓ,

1. Sample $\boldsymbol{\theta}_p^{(\ell)}$ from its conditional posterior distribution given the current data,

$$p(\boldsymbol{\theta}_p | y_1^{(\ell)}, \ldots, y_{p-1}^{(\ell)}, y_p^{(\ell-1)}, \ldots, y_P^{(\ell-1)}, \mathbf{X}_p). \tag{7.2}$$

2. For each $i = 1, \ldots, n_p^I$ subjects with $M_{ip} = 1$, draw a value $y_{ip}^{I(\ell)}$ from the posterior predictive distribution for the missing data conditional on the most recently drawn values,

$$g_p(y_{ip}^I | y_{i1}^{(\ell)}, \ldots, y_{ip-1}^{(\ell)}, y_{ip+1}^{(\ell-1)}, \ldots, y_{iP}^{(\ell-1)}, \boldsymbol{\theta}_p^{(\ell)}, \mathbf{x_{ip}}). \tag{7.3}$$

3. If $\boldsymbol{\phi}_p$ contains additional parameters not contained in $\boldsymbol{\theta}_p$, sample $\boldsymbol{\phi}_p^{(\ell)}$ from its conditional posterior distribution,

$$p(\boldsymbol{\phi}_p | y_1^{(\ell)}, \ldots, y_{p-1}^{(\ell)}, y_p^{(\ell-1)}, \ldots, y_P^{(\ell-1)}, \mathbf{X}_p). \tag{7.4}$$

4. For each $i = 1, \ldots, n_p^E$ subjects with $E_{ip} = 1$, draw $y_{ip}^{E(\ell)}$ from the editing model for the inconsistent data conditional on the most recently drawn values,

$$h_p(y_{ip}^E | y_{i1}^{(\ell)}, \ldots, y_{ip-1}^{(\ell)}, y_{ip+1}^{(\ell-1)}, \ldots, y_{iP}^{(\ell-1)}, \boldsymbol{\phi}_p^{(\ell)}, \mathbf{x_{ip}}). \tag{7.5}$$

5. The complete imputed and edited true data vector $y_p^{(\ell)} = (y_{1p}^{(\ell)}, \ldots, y_{np}^{(\ell)})'$ is then created as

$$z_{ip}^{(\ell)} = \begin{cases} z_{ip}^{ok} & \text{if } M_{ip} = 0 \cap E_{ip} = 0 \\ y_{ip}^{E(\ell)} & \text{if } M_{ip} = 0 \cap E_{ip} = 1 \\ y_{ip}^{I(\ell)} & \text{if } M_{ip} = 1 \end{cases} \tag{7.6}$$

for $i = 1, \ldots, N$.

7.2 Example 1: Modeling Inconsistent Multivariate Responses Using SyBRMICE

This example uses SyBRMICE to jointly impute missing data and edit the inconsistent multivariate responses (IMV) for knowledge (KNOW) and utilization (UTIL) of the Condom Availability Program (CAP). The IMV models from Chapter 6 fit the regression model (6.6a) on utilization for L iterations, then fit model (6.6b) on knowledge for an additional L iterations. The IMV editing step then occurred after both models were fit. The SyBRMICE procedure differs by fitting both utilization and knowledge models, and performing an imputation and editing step all within a single iteration. Draws of all unknown parameters from one iteration are fed into the next.

The priors and starting values are defined as in section 6.3.2. The data set drops cases with missing predictor variables and uses the deterministically edited values for gender and birthplace. The second example in this chapter provides an example where all available data is used: no records are excluded from the analysis.

7.2.1 Data

This example uses data from $N = 22,251$ surveys from high school students. Table 7.1 shows that $n_1^{mis} = 139 + 75 = 214$ are missing a response for the utilization question (Y_1) only, $n_2^{mis} = 80 + 21 = 101$ are missing a response for the knowledge question (Y_2) only, and 2414 are missing both. The missing data indicator $M_{i1} = 1$ if y_{i1} is missing, and $M_{i2} = 1$ if y_{i2} is missing for all $i = 1, \ldots, N$. Of those that provided data on both variables, $n_1^{err} = 673$ (3.5%) report the IMV response combination ($E_{i1} = 1$) of not knowing that someone on campus gives out condoms, but that they have received condoms from this person before.

The $q = 9$ fully observed predictor variables include an intercept, ethnicity (Hispanic, African American, Other), standardized age (AGE), presence of a school based health center (SBHC), indicator of being in the intervention condition (INTERV), being female (FEMALE), being born in the United States (US), and an indicator of ever having had sexual

		Knowledge (KNOW)			
		0	1	Missing	Total
Utilization	0	**10802**	**5395**	**80**	16277
	1	**673**	**2652**	**21**	3346
	Missing	**139**	**75**	**2414**	2628
	Total	11614	8122	2515	22251

Table 7.1: Cross-tabulation of utilization by knowledge of the CAP among high school students.

intercourse (SEXP). So

$$\mathbf{X}_1 = (1,\ \text{AA},\ \text{OTH},\ \text{AGE},\ \text{SBHC},\ \text{INTERV},\ \text{FEMALE},\ \text{US},\ \text{SEXP})'$$

is the $N \times q$ data matrix used to model Y_1 and $\mathbf{X}_2$ is the $N_0 \times q$ data matrix used to model Y_2, where $N_0 = 16,277$ rows with $y_{i1} = 0$. The regression, imputation, and editing models specific for Y_1 and Y_2 are defined first, followed by a description of how they are used in the SyBRMICE algorithm.

7.2.2 Model Definitions

Regression models. The probability ϕ_{i1} that student i has gotten a condom from the CAP is modeled using (7.7a), and conditional on not utilizing the CAP ($y_{i1} = 0$), the probability ϕ_{i2} of student i knowing they can get condoms from someone on campus is modeled using (7.7b)

$$\begin{aligned} y_{i1} &\sim \text{Bernoulli}(\phi_{i1}) \\ \phi_{i1} &= \Phi(\mathbf{x_{i1}}'\boldsymbol{\theta}_1) \end{aligned} \tag{7.7a}$$

$$\begin{aligned} (y_{i2}|y_{i1} = 1) &= 1 \text{ with probability } 1 \\ (y_{i2}|y_{i1} = 0) &\sim \text{Bernoulli}(\phi_{i2}) \\ \phi_{i2} &= \Phi(\mathbf{x_{i2}}'\boldsymbol{\theta}_2). \end{aligned} \tag{7.7b}$$

Multivariate normal priors are placed on the regression coefficients $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$

$$\boldsymbol{\theta}_1 \sim \mathcal{N}\left((-2.5, 0, 0, 0, 1, 0, 0, 0, 1)', \frac{19623}{25}(W_1' W_1)^{-1}\right), \tag{7.8a}$$

$$\boldsymbol{\theta}_2 \sim \mathcal{N}\left((-1.5, 0, 0, 0, 1, 0, 0, 0, 0)', \frac{16197}{25}(W_2' W_2)^{-1}\right), \tag{7.8b}$$

where the prior variances are calculated using only the 19,623 rows with observed Y_1 and 16,197 rows with observed $Y_2|Y_1 = 0$. The data matrix W_1 contains the rows i of $\mathbf{X}_1$ where y_{i1} was observed and W_2 contains the rows i of $\mathbf{X}_2$ where y_{i2} was observed and $y_{i1}^{obs} = 0$. The prior means are set based on observing trends in early test runs of this model using non-informative priors. The scaling factor $c = 25$ was determined by varying c in these test runs and choosing the one that produced a prior variance larger than the resulting posterior variance for each regression parameter, yet not too diffuse to be uninformative.

Imputation model. The imputation models for Y_1 and Y_2 contain logical constraints so that imputed values do not result in an inconsistent combination. Consider draws $\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$ from the posterior densities $p(\boldsymbol{\theta}_1|y_1, \mathbf{X}_1)$ and $p(\boldsymbol{\theta}_2|y_2, y_1 = 0, \mathbf{X}_2)$ respectively at the current iteration of the algorithm. Then for all $i = 1, \ldots, n_1^{mis}$ with $M_{i1} = 1$, imputed values y_{i1}^I are drawn from

$$y_{i1}^I|\boldsymbol{\theta}_1^*, y_{i2}, \mathbf{x_{i1}} \begin{cases} = 0 \text{ with probability } 1 & \text{if } y_{i2} = 0 \\ \sim \text{Bernoulli}(\Phi(\mathbf{x_{i1}'}\boldsymbol{\theta}_1)) & \text{if } y_{i2} = 1 \end{cases} \tag{7.9a}$$

and for all $i = 1, \ldots, n_2^{mis}$ with $M_{i2} = 1$, imputed values y_{i2}^I are drawn from

$$y_{i2}^I|\boldsymbol{\theta}_2^*, y_{i1}, \mathbf{x_{i2}} \begin{cases} \sim \text{Bernoulli}(\Phi(\mathbf{x_{i2}'}\boldsymbol{\theta}_2)) & \text{if } y_{i1} = 0 \\ = 1 \text{ with probability } 1 & \text{if } y_{i1} = 1. \end{cases} \tag{7.9b}$$

Since I am imputing prior to editing, I could ignore the logical constraints and allow the missing value to be imputed in an inconsistent manner. This would be similar to the *full edit* editing procedure where every record is stochastically edited regardless of their observed values. In this dissertation I primarily use the limited edit versions of the edit and imputation procedures. This is more aligned with the standard practice of imputing according to logical constraints as found in other multiple imputation software such as IVEWARE.

After the chains have been initialized, the imputation of y_1 depends on the most recently imputed value of y_2 and vice versa; the imputation of y_2 depends on the most recently imputed value of y_1. No further steps are necessary to accommodate when both y_1 and y_2 are missing because they will have been initialized prior to the first imputation. How starting values are set is discussed later.

Editing model. For each row with $E_{i1} = 1$, a pair of edited values (y_{i1}^E, y_{i2}^E) are created as follows. For row i at iteration (ℓ) the current value for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are used to calculate the imputation and editing cell probabilities.

$$\phi_{i1}^{(\ell)} = \Phi(\mathbf{x_{i1}'}\boldsymbol{\theta}_1^{(\ell)}), \tag{7.10a}$$

$$\phi_{i2}^{(\ell)} = \Phi(\mathbf{x_{i2}'}\boldsymbol{\theta}_2^{(\ell)}) \text{ when } y_{i1}^{(\ell)} = 0, \tag{7.10b}$$

$$p_{Ai}^{(\ell)} = (1 - \phi_{i1}^{(\ell)})(1 - \phi_{i2}^{(\ell)})\pi, \tag{7.10c}$$

$$p_{Bi}^{(\ell)} = (1 - \phi_{i1}^{(\ell)})\phi_{i2}^{(\ell)}\pi^2, \tag{7.10d}$$

$$p_{Di}^{(\ell)} = \phi_{i1}^{(\ell)}\pi, \tag{7.10e}$$

where $\pi = \frac{673}{2*22251} = 0.015$. A vector $R_i^{(\ell)}$ from a Multinomial distribution is then drawn

$$R_i^{(\ell)} \sim \text{ Multinomial}(1, p_{Ai}^{(\ell)}, p_{Bi}^{(\ell)}, p_{Di}^{(\ell)}), \tag{7.11a}$$

where the cell probabilities have been normalized prior to the draw, and the pair of edited values (y_{i1}^E, y_{i2}^E) is created as

$$(y_{i1}^E, y_{i2}^E)^{(\ell)} = \begin{cases} (0,0) & \text{if } R_i^{(\ell)} = (1,0,0) \\ (0,1) & \text{if } R_i^{(\ell)} = (0,1,0) \\ (1,1) & \text{if } R_i^{(\ell)} = (0,0,1). \end{cases} \tag{7.11b}$$

7.2.3 Algorithm

Initializing the chains. The starting values ($\ell = 0$) for all random variables (missing data and regression parameters) are created for each chain as follows.

1. Missing y_{i1} are drawn as a random Bernoulli(.328) variable if $y_{i2} = 1$ or missing. Set $y_{i1} = 0$ if $y_{i2} = 0$. The probability parameter .328 is calculated as the average observed value of y_{i1} across observations with $y_{i2} = 1$ or missing; $.328 = \frac{1}{2673} \sum_{i=1} y_{i1}$.

2. Missing y_{i2} are drawn as a random Bernoulli(.333) variable if $y_{i1}^{(0)} = 0$. Set $y_{i2} = 1$ if $y_{i1}^{(0)} = 1$. The probability parameter is calculated as the average value across observations with $y_{i1} = 0$ or missing; $.333 = \frac{1}{5740} \sum_{i=1} y_{i2}$.

3. Starting values for the elements of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are set as the prior mean plus or minus a random uniform deviate from [-1,1] to create diversity across chains.

The SyBRMICE procedure then proceeds as detailed in Algorithm 7.2.

7.2.4 Modeling Results

The final simulation was run for 4,200 iterations on each of 5 chains. After discarding the first 200 iterations per chain and keeping every 10^{th} iteration, the final posterior sample size was 2,000. Convergence diagnostic plots for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are located at the end of this chapter, appendix figures 7.3 and 7.4. These show that the chains were run long enough to achieve adequate mixing, convergence and approximately normal posterior densities.

Table 7.2 gives summary statistics for the posterior distribution of the probit regression parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$. Posterior mean estimates, standard deviations and 95% intervals are reported. Being older, having an SBHC, attending an intervention school, being US born and ever having sex are all associated with an increased probability of both knowing about and getting a condom from the CAP. Being female was associated with a lower probability of knowing about or getting a condom from the CAP. African-American students have lower, and other ethnicity have higher probabilities of both knowing about and utilizing the CAP

Algorithm 7.2 SyBRMICE Algorithm Applied to Inconsistent Multivariate Responses Between Knowledge and Utilization of the CAP.

Do from $\ell = 1$ to L

1. Model and update CAP utilization.
   (a) Fit model (7.7a) using all $y_1^{(\ell-1)}$, the complete and consistent data on utilization from the previous iteration.
   (b) Sample $\boldsymbol{\theta}_1^{(\ell)}$ from its posterior distribution (6.3).
   (c) For each observation with $M_{i1} = 1$, draw an imputed value $y_{i1}^{I(\ell)}$ from (7.9a).
2. Model and update CAP knowledge, conditional on the response for utilization.
   (a) Fit model (7.7a) using $y_{i2}^{(\ell-1)}$ the data on knowledge from the previous iteration conditional on the current value of utilization being zero, $y_{i1}^{(\ell)} = 0$.
   (b) Sample $\boldsymbol{\theta}_2^{(\ell)}$ from its posterior distribution (6.3).
   (c) For each observation with $M_{i2} = 1$ draw an imputed value $y_{i2}^{I(\ell)}$ from (7.9b).
3. For each observation with $E_{i2} = 1$, draw jointly edited values (y_{i1}^E, y_{i2}^E) using equation (7.11).
4. The observed, imputed and edited data are concatenated as in equation (7.6) to create complete and consistent vectors $y_1^{(\ell)}$ and $y_2^{(\ell)}$.

compared to Hispanic students.

I calculate the posterior distribution of $\Phi(\mathbf{c_p}'\boldsymbol{\theta}_p)$ to describe the posterior probabilities of knowing about, and getting a condom from the CAP for 2 fictitious students with covariate profiles c_1 and c_2. A foreign born, African-American male student who has had sex, attends a control high-school school without a school based health center and is 14.9 years old has a $\Phi\big((1,1,0,0,0,0,0,0,1)\boldsymbol{\theta}_1\big) = .06$ (95%PI .05, .08) probability of getting a condom from the CAP. Similarly a 14.9 year old Hispanic girl from Los Angeles attending an intervention high school that has a school based health center but has not had sex has a $\Phi\big((1,0,0,0,1,1,1,1,0)\boldsymbol{\theta}_2\big) = .52$ (.50, .53) probability of knowing about the CAP at her high school.

Figure 7.1(a) shows the posterior kernel density estimate (KDE) of the cell editing probabilities $p_{Ai}^{(\ell)}$, $p_{Bi}^{(\ell)}$ and $p_{Di}^{(\ell)}$, using all ℓ samples, defined in equations (7.10c) – (7.10e). This plot indicates that the probability p_{Di} of being edited into cell **D** is likely 0.2, or 0.6, and the probability p_{Ai} of being edited into cell **A** is near either 0.4 or 0.8. This bimodal distribution is driven by the SBHC; students attending schools with a school based health center have a higher posterior mean probability of being edited into cell **D** compared to students without SBHC's. The probability of an inconsistent record being edited into cell **B** is essentially 0. This means that virtually all inconsistent records are either edited into cell **A**, or cell **D**. Figure 7.1 (b) confirms this nearly perfect inverse relationship between p_{Ai} and p_{Di}. This displays data from the first 5 inconsistent records. The upper portion plots the editing cell probabilities $(p_{Ai}^{(\ell)}, p_{Di}^{(\ell)})$ against each other, with $y = 1 - x$ as a grey reference line. The lower plot shows the marginal posterior distribution of p_{Ai}, demonstrating that within a student, there is variability in the probability that the student would be edited into cell **A**. Including and accounting for this variability in the editing process is the core purpose of SyBRMICE. An edit must be made, but it should have a random component, and that added uncertainty should be carried through to the final modeling results.

7.2.5 MEMI Results

From the posterior samples, $M = 20$ random draws were used to generate the 20 MEMI data sets. Valid cell counts were generated on each MEMI data set, and the results combined using Rubin's Rules (3.1). The intervals are generated on the cell counts, then converted to percentages by dividing by N. The results are shown in Table 7.3 and show that 58.6% (57.9%, 59.3%) of students do not know about, nor have gotten a condom from the CAP; 27.8% (27.2%, 28.5%) know about, but have not gotten a condom from the CAP; and 13.6% (13.1%, 14.1%) know about and have gotten a condom from someone at their school.

Figure 7.2 provides a comparison of the marginal percentage of students who knew about, and who got a condom from the CAP as calculated across all observations on the complete case (CC) and combined MEMI data sets. The MEMI estimate of the percent of students getting a condom from the CAP is smaller than the percentage calculated on the complete case data; 13.6% (13.1%, 14.1%) vs. 17.0%. The estimates for the percent who knew about the CAP are slightly larger for the MEMI estimates compared to the complete case estimates; 41.5% (40.1%, 42.1%) vs. 40.1%.

The next section presents a complete and full SyBRMICE model, where multiple variables are subject to missing and/or inconsistent data, and the predictors used to model these variables are themselves subject to missing and/or inconsistent responses.

Parameter	Mean	SD	2.5%	97.5%	$p(\boldsymbol{\theta}_{qk} > 0 \mid Y)$
Utilization(Y_1)					
Intercept	−2.299	0.045	−2.384	−2.216	<0.001
African-American	−0.011	0.041	−0.092	0.071	0.386
Other	0.179	0.039	0.103	0.261	1.000
Standardized Age	0.026	0.013	0.000	0.051	0.976
SBHC	0.973	0.027	0.919	1.026	1.000
Intervention	0.204	0.025	0.154	0.253	1.000
Female	−0.164	0.024	−0.209	−0.115	<0.001
US Born	0.150	0.031	0.088	0.210	1.000
Sexually Experienced	0.755	0.027	0.705	0.808	1.000
Knowledge | *No Utilization*$(Y_2 \mid Y_1 = 0)$					
Intercept	−1.150	0.034	−1.215	−1.083	<0.001
African-American	−0.065	0.034	−0.133	0.002	0.028
Other	0.163	0.033	0.097	0.229	1.000
Standardized Age	0.039	0.011	0.016	0.060	1.000
SBHC	0.795	0.020	0.756	0.834	1.000
Intervention	0.289	0.021	0.248	0.330	1.000
Female	−0.087	0.022	−0.130	−0.045	<0.001
US Born	0.190	0.026	0.138	0.241	1.000
Sexually Experienced	0.093	0.022	0.049	0.137	1.000

Table 7.2: Summary of the probit regression parameter posteriors for the SyBRMICE CAP example of section 7.2.2.

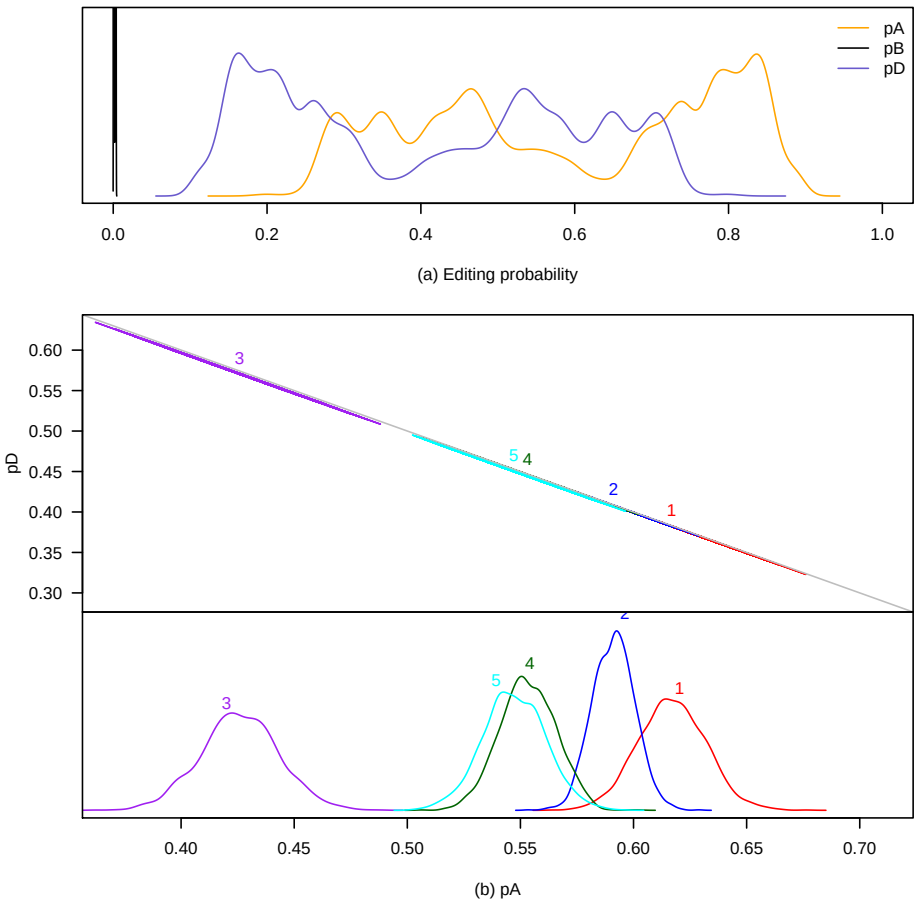


Figure 7.1: Posterior probability densities for the SyBRMICE CAP example. Subfigure (a) shows the distribution of cell editing probabilities for all iterations defined in equations (7.10c) – (7.10e). Subfigure (b) is the posterior marginal probability of being edited into cell **A** for $i = 1, \ldots, 5$, and the relationship between p_{Ai} and the probability of being edited into cell **D**, p_{Di}.

	Estimate	95% Interval	
%(No, No) **A**	58.55%	57.86%	59.25%
%(No, Yes) **B**	27.84%	27.21%	28.46%
%(Yes, Yes) **D**	13.61%	13.14%	14.08%

Table 7.3: Valid cell percent estimates with 95% intervals estimated using $M = 20$ SyBRMICE MEMI data sets and combined using Rubin's rules.

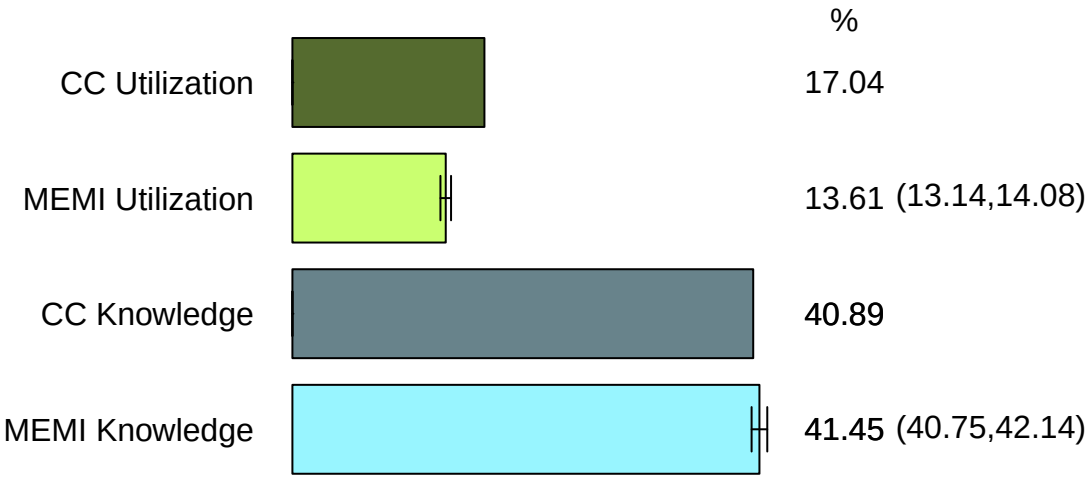


Figure 7.2: Comparing percent Knowledge and percent Utilization from complete cases (CC) to 20 SyBRMICE MEMI's.

7.3 Example 2: Combining Multiple Imputation and Editing Models Using SyBRMICE

This section combines all primary models presented in this dissertation into a single multivariate analysis that jointly imputes missing and edits inconsistent longitudinal reports of gender and birthplace, imputes and jointly edits inconsistent multivariate reports of CAP knowledge and utilization, and imputes and edits inconsistent monotone longitudinal reports of ever having sex. This example demonstrates the flexibility and customizability of the SyBRMICE procedure.

Let Y_1 = FEMALE, Y_2 = US, Y_3 = SEXP, Y_4 = UTIL, and Y_5 = KNOW be the five variables subject to missing and/or inconsistent data for this example. The first two variables, Y_1 and Y_2, are subject level variables so constant across all surveys, and the rest Y_3, Y_4 and Y_5 are observation level variables, meaning they can differ on each survey. Fully observed predictors are AGE_{ij}, SBHC_{ij}, INTERV_{ij}, WEAP_i, FIGHT_i, and ethnicity as two indicator variables AA_i, OTH_i, using Hispanic as the reference group. All variables were introduced earlier.

To create a cyclical framework so that the response level CAP variables feed back into the subject level gender and birthplace models, I introduce two new subject level variables y_4^*, y_5^* that are calculated as the average response of the corresponding utilization and knowledge variables y_{ij4}, y_{ij5} across all 9^{th} to 12^{th} grades. To maintain equivalence across subjects, these variables need to be averaged over the same number of surveys for each subject. This is the second reason for expanding the data set to create a fully balanced data as described in section 5.1. Not all models in this example will use these added variables. The unaltered data structure is the *response structure* while the structure that includes these additional rows is the *balanced structure*.

The general process is to first fit the IRM model of Chapter 4 to gender (Y_1) using y_4^* and y_5^* as predictors, impute missing values and edit the inconsistently reported gender. Next fit the IRM model on birthplace (Y_2) using the updated version of Y_1, y_4^*, and y_5^*. Missing values of Y_2 are imputed and inconsistently reported birthplace is edited.

The indicator of ever having sex (Y_3) is not modeled directly, but through the survival model on grade at first sex using the most recent versions of Y_1, Y_2, y_4^*, and y_5^* as covariates. The predicted grade at first sex for each student then is used to edit the inconsistent response patterns and consistently impute missing values for each subject i across all j surveys.

Utilization of the CAP (Y_4) is modeled with Y_1, Y_2, Y_3 as predictors, and missing values are imputed. Conditional on no utilization ($Y_4 = 0$), knowledge of the CAP (Y_5) is modeled using the same set of predictors as utilization, and missing values for Y_5 are imputed. Lastly the inconsistent pairs of (Y_4, Y_5) are jointly edited.

The next subsection provides details on each of the regression, editing and imputation models for all $Y_1, \ldots, Y_5$.

7.3.1 Model Definitions

Regression Model for Inconsistent Repeated Measures on Gender (Y_1). Using the IRM model from Chapter 4 on the response data structure, the probability ϕ_{i1} that student i is female is modeled as

$$\begin{aligned} z_{i1}|\phi_{i1} &\sim \text{Bernoulli}(\phi_{i1}) \\ \text{logit}(\phi_{i1}) &= \mathbf{w_{i1}'}\boldsymbol{\gamma}_1 \\ y_{ij1}|\pi_{ij1} &\sim \text{Bernoulli}(\pi_{ij1}) \\ \text{logit}(\pi_{ij1}) &= \mathbf{x_{ij1}'}\boldsymbol{\alpha}_1 z_{i1} + \mathbf{x_{ij1}'}\boldsymbol{\beta}_1(1 - z_{i1}), \end{aligned} \tag{7.12}$$

where $\mathbf{x_{ij1}} = (1,\ \text{AGE}_{ij})', \mathbf{w_{i1}} = (1,\ \text{WEAP}_i,\ \text{FIGHT}_i,\ y_{i4}^*,\ y_{i5}^*)'$, z_{i1} is the underlying latent value for Y_1, and y_{ij1} is the observed value for subject i at time j.

Multivariate normal priors are placed on the vectors of regression parameters $\boldsymbol{\alpha_1}, \boldsymbol{\beta_1}$ and $\boldsymbol{\gamma_1}$ with mean vectors $\mathbf{m}_{\boldsymbol{\alpha}1} = (5, 0)', \mathbf{m}_{\boldsymbol{\beta}1} = (-5, 0)'$, and $\mathbf{m}_{\boldsymbol{\gamma}1} = (0, -.5, -.5, 0, 0)'$. The prior mean vector $\mathbf{m}_{\boldsymbol{\gamma}1}$ is set to reflect the prior belief that females have a lower likelihood of carrying weapons or fighting (Centers for Disease Control and Prevention, 2010). The prior covariance matrices are calculated as $\mathbf{v_{\alpha 1}} = \mathbf{v_{\beta 1}} = N_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}/5$, and $\mathbf{v_{\gamma 1}} = n_1(\mathbf{W_1}'\mathbf{W_1})^{-1}/5$ where $\mathbf{X}_1$ has rows $\mathbf{x_{ij1}'}$, $\mathbf{W_1}$ has rows $\mathbf{w_{i1}'}$, and where N_1 and n_1 are the number of rows in $\mathbf{X}_1$ and $\mathbf{W_1}$ with no missing data respectively.

Regression Model for Inconsistent Repeated Measures on Birthplace (Y_2). The probability ϕ_{i2} that student i was born in the USA is modeled using the response data structure as

$$\begin{aligned} z_{i2}|\phi_{i2} &\sim \text{Bernoulli}(\phi_{i2}) \\ \text{logit}(\phi_{i2}) &= \mathbf{w_{i2}'}\boldsymbol{\gamma}_2 \\ y_{ijp}|\pi_{ijp} &\sim \text{Bernoulli}(\pi_{ijp}) \\ \text{logit}(\pi_{ijp}) &= \mathbf{x_{ij2}'}\boldsymbol{\alpha}_2 z_{i2} + \mathbf{x_{ij2}'}\boldsymbol{\beta}_2(1-z_{i2}), \end{aligned} \tag{7.13}$$

where $\boldsymbol{\alpha}_2, \boldsymbol{\beta}_2$ and $\boldsymbol{\gamma}_2$ are vectors of regression parameters, $\mathbf{x_{ij2}} = (\mathbf{1},\ \text{AGE}_{ij},\ z_{i1}^{(\ell)})'$ and $\mathbf{w_{i2}} = (1,\ \text{AA}_i,\ \text{OTH}_i,\ y_{i4}^*,\ y_{i5}^*)'$. The predictor $z_{i1}^{(\ell)}$ is the complete data vector of gender from the current iteration. How this is created will be described soon. The prior mean $\mathbf{m_{\gamma 2}} = (.85, 0, 0, 0, 0)'$ is set to reflect the prior belief that students of all ethnicities have a 70% chance of being born in the US, and that a priori knowledge and utilization of the CAP are not associated with birthplace. The prior covariance matrices are calculated as $\mathbf{v_{\alpha 2}} = \mathbf{v_{\beta 2}} = N_2(X_2'X_2)^{-1}/5$, and $\mathbf{v_{\gamma 2}} = n_2(W_2'W_2)^{-1}/5$, again where $\mathbf{X}_2$ has rows $\mathbf{x_{ij2}'}, \mathbf{W_2}$ has rows $\mathbf{w_{i2}'}$, and where N_2 and n_2 are the number of rows in $\mathbf{X}_2$ and $\mathbf{W_2}$ with no missing data respectively.

When the model for Y_p is hierarchical, as is the case with the IRM models on Y_1 and Y_2, sampling from the conditional posterior distribution of $p(\boldsymbol{\theta}_p|Y_p, X, \boldsymbol{\theta}_{-p})$ can consist of several Gibbs or M-H steps, where the M-H steps may use adaptive sampling. Due to this extra complexity of the model, values simulated from the IRM models take much longer to converge or achieve suitable mixing than do the samples drawn from the Gibbs sampling algorithm used in the IMV models. To accommodate this difference I run a mini-cycle of the IRM model within each larger cycle of the entire SyBRMICE algorithm. During each full iteration of the SyBRMICE algorithm, the IRM model loops R times on Y_1, the values of the unknown model parameters from the terminal iteration are saved. The complete and consistent version of Y_1, z_1 is then passed on to the next IRM model for Y_2 which then loops for R times, saving the values from the terminal iteration. The newly created z_1, z_2 are then passed to the next steps and the full cycle continues with the models for Y_3, Y_4 and Y_5 executed to complete one full SyBRMICE cycle. This is done in addition to adaptive

M-H sampling to get good estimates of the proposal variances for both IRM models prior to retaining samples in phase 2.

Imputation and Editing Inconsistent Longitudinal Responses of Y_1 and Y_2. Let $E_{i1} = 1$ if student i inconsistently reported their gender across surveys, and $E_{i2} = 1$ if they inconsistently reported their birthplace. The missing data indicators M_{i1} and M_{i2} are equal to 1 if student i did not report their gender or birthplace respectively on any survey and 0 otherwise. The individual missing responses of gender y_{ij1} or birthplace y_{ij2} are not imputed themselves.

Equation (7.14) simultaneously edits the inconsistent and imputes the missing data for gender or birthplace by creating $z_{ip}^{(\ell)}$, the current sampled value at iteration ℓ from the posterior of student i's true value. The true value is taken to be the reported value from the first non-missing response if the student reported this variable consistently, the modeled latent value if the student reported inconsistently, or drawn from a Bernoulli distribution using the student's covariate pattern if the variable was missing on every survey.

$$z_{ip}^{(\ell)} \quad \begin{cases} = y_{ijp} \text{ for consistent reports, when } E_{ip} = 0 \\ = z_{ip}^{(\ell)} \text{ for inconsistent reports, when } E_{ip} = 1, \\ \sim \text{Bernoulli}[\text{expit}(\mathbf{w_{ip}'}\boldsymbol{\gamma}_p^{(\ell)})], \text{ when } M_{ip} = 1, \end{cases} \tag{7.14}$$

where $p = 1$ or 2 for Y_1 and Y_2 respectively.

Regression Model for Inconsistent Monotone Longitudinal Reports on Ever had Sexual Intercourse (Y_3). Using the IML model of Chapter 5 on the balanced data structure, the grade t_i the student first reported ever having sex is modeled using an interval censored regression model

$$P(t_{1i} > t_i > t_{2i}) = \Phi(t_{2i}|\mathbf{x_{i3}'}\boldsymbol{\theta}_3, \sigma^2) - \Phi(t_{1i}|\mathbf{x_{i3}'}\boldsymbol{\theta}_3, \sigma^2), \tag{7.15}$$

where $\mathbf{x_{i3}} = (1,\ \mathrm{AA}_i,\ \mathrm{OTH}_i,\ \text{Age at entry}_i,\ z_{i1},\ z_{i2},\ y^*_{i4},\ y^*_{i5})'$. Priors are assigned to the regression parameters $\boldsymbol{\theta}_3$ and variance σ^2

$$\begin{aligned} \boldsymbol{\theta}_3 &\sim N(\mathbf{m}_3, \mathbf{v}_3), \\ \sigma^2 &\sim IG(.001, .001). \end{aligned} \tag{7.16}$$

The mean vector $\mathbf{m}_3 = (12, 0, 0, 0, 0, 0, -2, -1)'$ is set to reflect the belief that the average grade at first intercourse is 12th grade, and that both utilization and knowledge of the CAP are associated with having sex at earlier grades. All other variables are not expected to a priori be associated with grade at first sex. The prior covariance matrix is set to be $\mathbf{v}_3 = n_3(X_3'X_3)^{-1}/25$, where $\mathbf{X}_3$ has rows $\mathbf{x'_{i3}}$ and $n_3 = n$, the total number of students in the sample.

Imputation and Editing models for Ever had Sex. Missing data indicators $M_{ij3} = 1$ are assigned if y_{ij3} was not observed. An inconsistent monotone longitudinal response occurs if a 0 was observed after a 1 was reported, so $E_{i3} = 1$ if $y_{ij3} = 1$ and $y_{ij'3} = 0$ when $j < j'$. Consider draws $\boldsymbol{\theta}^*_3$ and σ^{2*} from their posterior distributions given the data. The predicted grade of first sex t^*_i, is drawn from a truncated normal distribution (5.4) with mean $\mathbf{x'_{i3}}\boldsymbol{\theta}^*_3$ and variance σ^{2*}, and is used to edit and impute the response level values for y_{ij3} as follows. Missing and inconsistent data are imputed and edited simultaneously and the sampled complete data vector $y^{(\ell)}_{ij3}$ is

$$y^{(\ell)}_{ij3} \begin{cases} = y_{ij3} & \text{if } M_{ij3} = 0 \text{ and } E_{i3} = 0 \\ = 0 & \text{if } (M_{ij3} = 1 \text{ or } E_{i3} = 1) \text{ and } \text{grade}_{ij} < t^*_i \\ = 1 & \text{if } (M_{ij3} = 1 \text{ or } E_{i3} = 1) \text{ and } \text{grade}_{ij} \geq t^*_i, \end{cases} \tag{7.17}$$

The `MCMCglmm` function also uses an adaptive M-H sampling algorithm, so similar to the IRM models this survival model is run for an additional R_x iterations per full cycle of the SyBRMICE algorithm with the values from the terminal iteration passed on as starting values in the next cycle. During phase 2 all adaptive sampling is disabled, ensuring a constant proposal distribution for the remainder of the simulation.

Model for Inconsistent Multivariate Responses between Utilization (Y_4) **and Knowledge** (Y_5) **of the CAP.** Using the SyBRMICE IMV models described in Example 1, the model of utilization and knowledge of the CAP is fit to the N_c 9$^{\text{th}}$ through 12$^{\text{th}}$ grade surveys from the balanced data structure. Whether a student has gotten a condom from the CAP is modeled as

$$\begin{aligned} y_{ij4}|\phi_{ij4} &\sim \text{Bernoulli}(\phi_{ij4}) \\ \phi_{ij4} &= \Phi(\mathbf{x}_{\mathbf{ij4}}^{'}\boldsymbol{\theta}_4) \end{aligned} \tag{7.18}$$

where

$$\begin{aligned} \mathbf{x_{ij4}} &= (1,\ \text{AA}_i,\ \text{OTH}_i,\ \text{AGE}_{ij},\ \text{SBHC}_{ij},\ \text{INTERV}_{ij}, z_{i1}, z_{i2}, y_{ij3})^{'} \\ \boldsymbol{\theta}_4 &\sim \mathcal{N}\left((-2.5, 0, 0, 0, 1, 0, 0, 0, 1)^{'}, \frac{N_c}{25}(\mathbf{X}_4^{'}\mathbf{X}_4)^{-1}\right), \end{aligned}$$

where $\mathbf{X}_4$ has rows $\mathbf{x}_{\mathbf{ij4}}^{'}$.

Imputing CAP Utilization. If y_{ij4} is missing then M_{ij4} is set to 1 and equation (7.9a) is used to sample imputed values y_{ij4}^I for the missing data.

Model CAP Knowledge. Using the balanced data structure on high school students, and conditional on not utilizing the CAP (from either the observed consistent data, or the most recently imputed data) the probability a student knows they can get condoms from someone on campus, ϕ_{ij5}, is modeled as

$$\begin{aligned} (y_{ij5}|Z_{i4} = 1) &= 1 \text{ with probability } 1 \\ (y_{ij5}|Z_{i4} = 0, \phi_{ij5}) &\sim \text{Bernoulli}(\phi_{ij5}) \\ \phi_{i5} &= \Phi(\mathbf{x}_{\mathbf{ij5}}^{'}\boldsymbol{\theta}_5). \end{aligned} \tag{7.19}$$

where

$$\begin{aligned} \mathbf{x_{ij5}} &= (1,\ \text{AA}_i,\ \text{OTH}_i,\ \text{AGE}_{ij},\ \text{SBHC}_{ij},\ \text{INTERV}_{ij},\ z_{i1},\ z_{i2},\ y_{ij3})^{'} \\ \boldsymbol{\theta}_5 &\sim \mathcal{N}\left((-1.5, 0, 0, 0, 1, 0, 0, 0, 0)^{'}, \frac{N_c}{25}(\mathbf{X}_5^{'}\mathbf{X}_5)^{-1}\right), \end{aligned}$$

where $\mathbf{X}_5$ has rows $\mathbf{x}_{\mathbf{ij5}}^{'}$.

Imputing CAP Knowledge. If y_{ij5} is missing then M_{ij5} is set to 1 and equation (7.9b) is used to sample imputed values $y_{ij5}^{I(\ell)}$ for the missing data.

Joint Editing of Utilization and Knowledge. The inconsistent data indicator E_{ij4} is set to 1 if student i reported the inconsistent combination of no knowledge but utilization of the CAP on survey j and 0 otherwise. When $E_{ij4} = 1$, a pair of edited values (y_{ij4}^E, y_{ij5}^E) are drawn using the IMV *limited edit* procedure described in equation (6.8). After each of these editing or imputation procedures is performed, the complete and consistent variables y_{ij4} and y_{ij5} are created using equation (7.11) and the aggregated values y_4^* and y_5^* are calculated.

7.3.2 Algorithm

Starting the chains The starting values ($\ell = 0$) for the missing data and regression parameters are created for each of s chains:

1. For missing or inconsistent Y_1, Y_2, starting values for the modeled latent variables z_{i1}, and z_{i2} are drawn from a Bernoulli distribution with probability parameters equal to the average of that variable across subjects without missing or inconsistent data.

2. Starting values for the missing y_{ij3} are created using equation (5.6).

3. Missing values for knowledge and utilization are set as in section 7.2.3.

4. Starting values $y_4^{*(0)}$ and $y_5^{*(0)}$ are calculated from $y_{ij4}^{(0)}$ and $y_{ij5}^{(0)}$, the observed or imputed values for y_4 and y_5.

5. Starting values for all regression coefficients are set equal to the prior mean $\pm$ a random uniform deviate from [-1,1] to create diversity across chains.

Since most predictors do not change across iterations, they can be thought of as having starting values that are constant. I use this concept to shorten notation of the set of predictor variables. For example consider the predictors for $Y_4, \mathbf{x}_4$ = (1, AA, OTH, AGE, SBHC, INTERV, z_1, z_2, $y_3)'$. The values for ethnicity, age, SBHC and intervention condition do

not change across iterations, only true gender, birthplace, and ever sex (z_1, z_2, y_3) update. I define $\mathbf{x}_4^{(0)} = (1, \text{AA}, \text{OTH}, \text{AGE}, \text{SBHC}, \text{INTERV})'$. Then at iteration ℓ, $\mathbf{x}_4^{(\ell)} = (\mathbf{x}_4^{(0)}, z_1^{(\ell)}, z_2^{(\ell)}, y_3^{(\ell)})'$. Using this notation the process follows algorithm 7.1, where the first iteration uses the starting values $y_{ijp}^{(0)}$ and subsequent iterations use the complete data vectors $z_p^{(\ell-1)}$ for p=1, 2 and $y_p^{(\ell-1)}$ for p=3, 4, 5 from the previous iteration.

At iteration ℓ the Sequential Bayesian Regression for Multiple Imputation and Conditional Editing (SyBRMICE) algorithm is

1. Model and update the student's true gender z_1.
   (a) Construct the predictor matrix $\mathbf{w}_1^{(\ell)} = (\mathbf{w}_1^{(0)}, y_4^{*(\ell-1)}, y_5^{*(\ell-1)})$.
   (b) Fit the IRM model (7.12) to $z_1^{(\ell-1)}$ with predictors $\mathbf{w}_1^{(\ell)}$, and $\mathbf{x}_1$.
   (c) Sample $\boldsymbol{\gamma}_1^{(\ell)}$ from its posterior distribution.
   (d) For each subject with $M_{i1} = 1$ or $E_{i1} = 1$, impute missing and edit inconsistent gender y_{i1} using equation (7.14) and $\boldsymbol{\gamma}_1^{(\ell)}$ to create $z_{i1}^{(\ell)}$.
2. Model and update the student's true birthplace z_2.
   (a) Construct the predictor matrices $\mathbf{w}_2^{(\ell)} = (\mathbf{w}_2^{(0)}, y_4^{*(\ell-1)}, y_5^{*(\ell-1)})$ and $\mathbf{x}_2^{(\ell)} = (\mathbf{x}_2^{(0)}, z_1^{(\ell)})$.
   (b) Fit the IRM model (7.13) to $z_2^{(\ell-1)}$ using predictors $\mathbf{w}_2^{(\ell)}$, and $\mathbf{x}_\mathbf{2}^{(\ell)}$.
   (c) Sample $\boldsymbol{\gamma}_2^{(\ell)}$ from its posterior distribution.
   (d) For each subject with $M_{i2} = 1$ or $E_{i2} = 1$, impute missing and edit inconsistent birthplace y_{i2} using equation (7.14) and $\boldsymbol{\gamma}_2^{(\ell)}$ to create $z_{i2}^{(\ell)}$.
3. Model and update lifetime sexual experience y_{ij3} using the balanced data structure.
   (a) Construct the predictor matrix $\mathbf{x}_3^{(\ell)} = (\mathbf{x}_1, z_1^{(\ell)}, z_2^{(\ell)})$.
   (b) Fit model (7.15) using predictors $\mathbf{x}_3^{(\ell)}$.
   (c) Sample $\boldsymbol{\theta}_3^{(\ell)}$ from its posterior distribution and sample the predicted grade at first sex $t_i^{(\ell)}$ only for those missing or need editing.

(d) For each observation with $M_{ij3} = 1$ or $E_{i3} = 1$, impute or edit y_{ij3} using equation (7.17) to create $y_{ij3}^{(\ell)}$.

4. Model and update CAP utilization y_{ij4} using the balanced data structure for 9^{th} – 12^{th} grades.

   (a) Create the predictor matrix $\mathbf{x}_4^{(\ell)} = (\mathbf{x}_4, z_1^{(\ell)}, z_2^{(\ell)}, y_3^{(\ell)})$.

   (b) Fit model (6.6a) to $y_{ij4}^{(\ell-1)}$ using predictors $\mathbf{x}_4^{(\ell)}$.

   (c) Sample $\boldsymbol{\theta}_4^{(\ell)}$ from its posterior distribution.

   (d) For each observation with $M_{ij4} = 1$, impute y_{ij4} using equation (7.9a), $\boldsymbol{\theta}_4^{(\ell)}$ and $y_{ij5}^{(\ell-1)}$ to create $y_{ij4}^{I(\ell)}$.

5. Model and update CAP knowledge y_{ij5} using the balanced data structure for 9^{th} – 12^{th} grades and given no utilization.

   (a) Construct predictor matrix $\mathbf{x}_5^{(\ell)} = (\mathbf{x}_5, z_1^{(\ell)}, z_2^{(\ell)}, y_3^{(\ell)})$.

   (b) Fit the IMV model (6.6b) to $y_{ij5}^{(\ell-1)}$ on the observations with $y_{ij4} = 0$ and using predictors $\mathbf{x}_5^{(\ell)}$.

   (c) Sample $\boldsymbol{\theta}_5^{(\ell)}$ from its posterior distribution.

   (d) For each observation with $M_{ij5} = 1$, impute y_{ij5} using equation (7.9b), $\boldsymbol{\theta}_5^{(\ell)}$ and $y_{ij4}^{(\ell)}$ to create $y_{ij5}^{I(\ell)}$.

6. For each observation with $E_4 = 1$, jointly edit (y_{ij4}, y_{ij5}) using algorithm (6.8), $\boldsymbol{\theta}_4^{(\ell)}$ and $\boldsymbol{\theta}_5^{(\ell)}$ to create $(y_{ij4}^{E(\ell)}, y_{ij5}^{E(\ell)})$.

   (a) Create $y_{ij4}^{(\ell)}$ and $y_{ij5}^{(\ell)}$ using equation (7.11).

   (b) Calculate $y_4^{*(\ell)}$ and $y_5^{*(\ell)}$ for use in Step 1.

This cycle repeats for $\ell = 2, \ldots, L$ iterations on each of s chains using appropriate burn-in and thinning values.

			Students	Surveys
	Total	N	26,851	36,573
	High School	N	19,079	23,796
Gender (Y_1)	Missing	n_1^{mis}	245	-
	Inconsistent	n_1^{err}	57	-
Birthplace (Y_2)	Missing	n_2^{mis}	194	-
	Inconsistent	n_2^{err}	170	-
Ever had Sex (Y_3)	Missing	n_3^{mis}	-	1,971
	Inconsistent	n_3^{err}	248	-
Utilization (Y_4)	Missing	n_4^{mis}		3,047
Knowledge (Y_5)	Missing	n_5^{mis}		2,898
IMV (Y_4, Y_5)	Inconsistent	n_4^{err}		730

Table 7.4: Sample size and amount of missing and erroneous data for the second SyBRMICE example. Total is middle and high school combined.

7.3.3 Modeling Results

Table 7.4 shows the sample sizes and amount of missing and inconsistent values for the outcome variables in this example. This example uses data from 36,573 surveys from 26,851 students in both middle and high school. Of these $n_1^{mis} = 245$ are missing all responses for gender, with $n_1^{err} = 57$ reporting their gender inconsistently. Another $n_2^{mis} = 194$ students are missing all responses for birthplace, with $n_2^{err} = 170$ reporting their birthplace inconsistently. There are $n_3^{mis} = 1,971$ surveys missing data on if they have ever had sexual intercourse, and $n_3^{err} = 248$ reported an inconsistent monotone response pattern. Of 23,796 surveys from high school students, $n_3^{err} = 730$ surveys have an inconsistent combination of utilization and knowledge of the CAP.

Then a variety of simulations were run as testing blocks to determine the thinning, R and R_x values necessary to achieve convergence, suitable mixing, and low autocorrelation.

Phase 1 was run on 5 chains for 500 iterations per chain with $R = 30, R_x = 100$ and a thin of 10. This was done to assess computation time and to allow the M-H proposal variances to adapt.

Phase 2 of the SyBRMICE algorithm consisted of $L = 2,000$ iterations with $R = 30, R_x = 100$ and a thin of 10 on each of $s = 5$ chains producing a final posterior sample size of $\tilde{n} = 1,000$. Convergence diagnostic plots are displayed in Chapter Appendix figures 7.5 – 7.9 (a) – (d), and indicate good mixing, low autocorrelation and smooth posterior densities. Tables 7.5 – 7.10 give a summary of the posterior distribution for the modeled parameters. For the IRM models on Y_1 and Y_2, probabilities (p in *italics*) for the intercepts and Odds Ratios (OR) for all other covariates with corresponding 95% intervals are reported. The parameter estimates for the probit regression models on Y_4 and Y_5 remain untransformed.

The point estimates and 95% intervals for the regression parameters for the ILR gender model are nearly identical to those in Chapter 4. Carrying a weapon in the 30 days prior to the survey date, fighting in the past year, knowing about and getting a condom from the CAP at any point during the study was associated with being Male. For both genders as age increases, the odds of reporting being female decreases.

The point estimates and 95% intervals for the regression parameters for the ILR model on birthplace are slightly different than those seen in Chapter 4. This is in part because the regression coefficients differ between the two models. Earlier the model of true birthplace was only modeled using ethnicity as three indicator variables with no intercept. This model is reference coded, where Hispanic is the reference group.

The probability of a Hispanic/Latino student who never reported knowing about or getting a condom from the CAP being born in the US is .736 (.722, .749). African-American and Other ethnicity students have 10.1 (7.8, 13.2) and .469 (.427, .517) respectively times the odds of Hispanics of being born in the US. The large odds ratio that an African American student is born in the US compared to a Hispanic/Latino student is comparable with the observed OR of 7.7 as calculated on the complete cases as displayed in Table 7.7.

Reporting utilizing the CAP anytime during the study is not significantly associated

Parameter	Est	SD	OR/p	2.5%	97.5%	$p(\boldsymbol{\theta}_{1k} > 0 \mid Y)$
Model of true female gender						
Intercept	−0.10	0.04	*0.475*	*0.457*	*0.492*	0.003
Carried a weapon	−0.95	0.05	0.385	0.347	0.425	<0.001
Fought past 12mo	−0.27	0.03	0.763	0.719	0.811	<0.001
Average utilization	−0.76	0.16	0.467	0.344	0.631	<0.001
Average knowledge	−0.38	0.07	0.682	0.593	0.780	<0.001
Models for reporting female gender						
Females						
Intercept	5.38	0.18	*0.995*	*0.994*	*0.997*	1.000
Age (standardized)	−0.34	0.20	0.711	0.487	1.022	0.036
Males						
Intercept	−5.79	0.23	*0.003*	*0.002*	*0.005*	<0.001
Age (standardized)	−0.33	0.21	0.719	0.471	1.065	0.062

Table 7.5: Summary of the posterior distributions for the regression coefficients used in the model of true gender. Odds Ratios (OR) or probabilities (p in *italics*) with corresponding 95% posterior intervals are included.

with place of birth, but students who reported knowing that they could get a condom from someone on their high-school campus at least once during the study had 2.45 (2.02, 3.02) times the odds of being born in the US than those who did not know about the CAP. The probability of accurately reporting birthplace decreases as age increases for those born in the US, and increases with age for those born elsewhere. Females are more likely than males to correctly report birthplace.

The point estimates and 95% intervals for the regression parameters for the IML response model of lifetime sexual experience are similar to those in Chapter 5. A Hispanic male student who was born in the US and entered the Project Connect study at 14.9 years old (the average) is expected on average to have engaged in sexual intercourse by the $10.08 + .019(14.9) =$

Parameter	Est	SD	OR/p	2.5%	97.5%	$p(\boldsymbol{\theta}_{1k} > 0 \mid Y)$
Model of true US born						
Intercept	1.025	0.036	*0.736*	*0.722*	*0.749*	1.000
African-American	2.316	0.135	10.14	7.837	13.188	1.000
Other	−0.757	0.048	0.469	0.427	0.517	<0.001
Average utilization	0.401	0.207	1.494	1.008	2.242	0.976
Average knowledge	0.902	0.084	2.465	2.045	2.870	1.000
Models for reporting US birthplace						
US Born						
Intercept	4.645	0.183	*0.990*	*0.987*	*0.993*	1.000
Age (standardized)	−1.374	0.160	0.255	0.186	0.347	<0.001
Female	0.401	0.166	1.493	1.091	2.096	0.994
Foreign Born						
Intercept	−3.043	0.171	*0.046*	*0.033*	*0.061*	<0.001
Age (standardized)	−0.931	0.113	0.394	0.316	0.491	<0.001
Female	−0.793	0.240	0.453	0.281	0.721	0.002

Table 7.6: Summary of the posterior distributions for the regression coefficients used in the model of true birthplace. Odds Ratios (OR) or probabilities (p in *italics*) with corresponding 95% posterior intervals are included.

10.4^{th} grade.

Many of the point estimates and 95% intervals for the regression parameters for the IMV response model on CAP utilization in Table 7.9 are similar to those seen in Table 7.2. Now African-American students are seen to have a significantly lower probability of knowing about the CAP compared to Hispanics, and as age increases the probability that the student will know about the CAP also significantly increases.

Some difference is seen between the two examples in the estimates and 95% intervals for the regression parameters on CAP Knowledge given no utilization. Specifically the estimate

	Foreign Born	US Born
Hispanic	4408	15895
African-American	111	3085

Table 7.7: Cross tabulation of ethnicity and birthplace.

Parameter	Est	SD	2.5%	97.5%	$p(\boldsymbol{\theta}_{qk} > 0 \mid Y)$
Intercept	10.080	0.269	9.558	10.653	1.000
African-American	−0.858	0.076	−1.002	−0.706	<0.001
Other	0.645	0.075	0.502	0.798	1.000
Age at Entry	0.019	0.017	−0.016	0.051	0.867
Female	0.882	0.045	0.794	0.969	1.000
US Born	−0.023	0.065	−0.147	0.106	0.378
σ	2.696	0.048	2.603	2.785	1.000

Table 7.8: Summary of the posterior distributions for the regression coefficients and variance parameter Sigma used in the model of grade at first sex.

for African-American did not change by much (was -.065 (-.133, .002)), but the parameter is no longer significant with a p-value of .064 compared to .028 from Example 1. While the difference in p-values itself is not significant, researchers who hold a hard line on anything less than .05 being significant, these are different results. The estimate for ever having sex is not significant with a p-value of .56, when in Example 1 this estimate was significantly greater than 0 (p-value > .999).

I reconsider the same covariate profiles as examined earlier and compare the fitted probabilities of knowing about and utilizing the CAP from this model to the one in Example 1. A foreign born, African-American male student who has had sex, attends a control high-school school without a school based health center and is 14.9 years old (average age of HS students) has a .01 (.007, .012) probability of getting a condom from the CAP. This is lower than the

Parameter	Est	SD	2.5%	97.5%	$p(\boldsymbol{\theta}_{qk} > 0 \mid Y)$
Intercept	-2.601	0.033	-2.666	-2.538	<0.001
African-American	-0.099	0.031	-0.161	-0.038	<0.001
Other	0.077	0.029	0.017	0.131	0.993
Age (standardized)	0.791	0.021	0.750	0.835	1.000
SBHC	0.789	0.021	0.748	0.830	1.000
Intervention	0.023	0.020	-0.016	0.061	0.880
Female	-0.137	0.019	-0.173	-0.100	<0.001
US Born	0.132	0.024	0.088	0.182	1.000
Ever had sex	0.358	0.018	0.324	0.391	1.000

Table 7.9: Summary of the posterior distributions for the regression coefficients used in the model of true utilization of the CAP.

estimated probability of .063 from Example 1. Similarly a 14.9 year old Hispanic girl from Los Angeles attending an intervention high school that has a school based health center but has not had sex has a .49 (.47, .51) probability of knowing about the CAP at her high school. This is modestly lower than the estimated probability of .53 from Example 1.

7.3.4 Multiple Editing and Multiple Imputation Results

To create the Multiply Edited and Multiple Imputed MEMI data sets, 20 samples were drawn from the posterior samples of all regression coefficients and imputed/edited values. This was done by randomly sampling 20 values from 1 to L and pulling all model results for those iterations.

Table 7.11 – 7.14 show the original response pattern, and the subsequent imputed and/or edited data for a sample of observations subject to missing and/or inconsistent data. These tables show how SyBRMICE can produce different edited or imputed values across MEMI data sets, demonstrating the between MEMI variance that is incurred as part of the stochastic

Parameter	Est	SD	2.5%	97.5%	$p(\boldsymbol{\theta}_{qk} > 0 \mid Y)$
Intercept	-1.060	0.031	-1.124	-1.000	<0.001
African-American	-0.048	0.032	-0.109	0.013	0.064
Other	0.149	0.032	0.085	0.210	1.000
Age (standardized)	0.071	0.016	0.037	0.102	1.000
SBHC	0.778	0.020	0.739	0.819	1.000
Intervention	0.183	0.020	0.143	0.223	1.000
Female	-0.087	0.019	-0.121	-0.050	<0.001
US Born	0.159	0.022	0.117	0.200	1.000
Ever had sex	0.002	0.011	-0.021	0.023	0.560

Table 7.10: Summary of the posterior distributions for the regression coefficients used in the model of true knowledge of the CAP given no utilization.

editing and imputation process.

ID	Raw	MEMI #1	MEMI #2	MEMI #3	MEMI #4	MEMI #5	Total F
1	FMFF	F	F	F	F	F	20/20
2	M.F	F	F	M	F	M	14/20
3	M.F	M	F	M	F	F	11/20
4	MF	M	M	M	M	F	6/20
5	FMM	M	M	M	M	M	0/20

Table 7.11: Edited and imputed values from five of MEMI data sets for five students with observed inconsistent longitudinal reports of gender. The total column shows how many times the student was edited to be Female. A value of . indicates a missing value.

ID	Raw	MEMI #1	MEMI #2	MEMI #3	MEMI #4	MEMI #5	Total US
6	FUs	US	US	US	F	US	18/20
7	UsUsF	F	US	US	US	US	15/20
8	FFUs	F	US	US	F	F	9/20
9	UsFF	US	F	F	F	US	8/30
10	UsFFUs	F	F	F	F	F	1/20

Table 7.12: Edited and imputed values from five MEMI data sets for five students with observed inconsistent longitudinal reports of birthplace. The total column shows how many times the student was edited to be born in the USA. F = Foreign Born, Us = US Born

ID	Grade	Raw `I1`	MEMI #1	MEMI #2	MEMI #3	MEMI #4	MEMI #5
11	10	0	1	1	0	0	0
	11	1	1	1	0	0	0
	12	0	1	1	0	1	1
12	10	.M	0	1	0	0	0
	11	.M	1	1	0	1	0
	12	.M	1	1	0	1	1
13	9	0	0	0	1	0	0
	10	.M	0	0	1	0	0
	11	1	1	0	1	1	0
	12	0	1	1	1	1	0
14	6	0	0	0	0	0	0
	7	1	0	0	0	1	0
	8	.M	1	0	0	1	0
	9	1	1	1	0	1	0

Table 7.13: Edited and imputed values from five MEMI data sets for five students with observed inconsistent longitudinal reports of ever having sex. A value of .M indicates a missing value, 0 = No sexual intercourse, 1 = Had sexual intercourse

	(Utilization, Knowledge)						Total times in cell		
ID	Raw	MEMI #1	MEMI #2	MEMI #3	MEMI #4	MEMI #5	A (0,0)	B (0,1)	D (1,1)
15	(1,0)	(0,0)	(0,0)	(1,1)	(0,0)	(0,0)	18	0	2
16	(1,0)	(0,0)	(1,1)	(1,1)	(0,0)	(0,1)	11	1	8
17	(.M, .M)	(0,0)	(0,1)	(0,0)	(0,1)	(1,1)	14	5	1
18	(.M, .M)	(1,1)	(0,1)	(0,1)	(1,1)	(0,0)	7	9	4
19	(0, .M)	(0,1)	(0,0)	(0,1)	(0,1)	(0,0)	10	10	0
20	(1, .M)	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)	0	0	20
21	(.M, 0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	20	0	0
22	(.M, 1)	(0,1)	(1,1)	(0,1)	(0,1)	(0,1)	0	16	4

Table 7.14: Edited and imputed values from five MEMI data sets for five students with observed inconsistent combinations of Utilization and Knowledge of the CAP. The total times the observation was reallocated into each of the three valid cells is include. A value of .M indicates a missing value, 0 = No, 1 = Yes.

Table 7.15 provides a full comparison of the different methods to estimate various proportions of interest in this dissertation, namely the % Female, % US Born, % Had sex, % Utilized the CAP, % Knew about the CAP given they reported no utilization, and the three valid cell percentages for the combination of utilization and knowledge. The complete case estimates for the percent female, and percent born in the USA, are calculated by averaging across students the first observed value from the students with consistent reporting patterns. These percentages then can be interpreted as the percent of students who are female, and the percent of students born in the USA. In contrast, the complete case estimates for ever had sex, knowledge and utilization are calculated by averaging across all complete and consistent survey responses. These percentages then represent the percent of surveys where a "Yes" response was provided to the variable of interest.

The top two rows in Table 7.15 display the percentages calculated from the complete cases, and the deterministically edited or imputed values used in the Project Connect data set. The bottom three rows are the percentages of interest with corresponding 95% Intervals calculated under all the different multiple editing and multiple imputation models presented in this dissertation. Across the board the point estimates differ at most by around 5% across calculation methods. For a sample of size 30,000, 5% represents a difference of 1,500 observations.

The % Female calculated under both the deterministic edit and the IRM model edit was 56.3%, but the IRM model provides a confidence interval of (55%, 57.5%). The SyBRMICE procedure estimated the % Female at 55.0% (54.5%, 55.5%) and the complete case analysis came in lowest with 54.3%. The US born complete case percentage is 79.1%, which is close to the SyBRMICE estimate of 80.6% (80.0%, 81.1%) with the IRM model resulting in a higher estimated percent born in the US at 82.0% (81.0%, 83.0%). Project Connect had no deterministic edit for an inconsistently reported birthplace.

The IML model for the % of surveys where a student reports ever having sexual intercourse is higher than both the complete case and SyBRMICE estimates at 34.8% (34.3%, 35.3%) for the IML model, 33.9% for the complete cases and 33.7% (33.1%, 34.2%) for the SyBRMICE model. Project Connect had no deterministic edit for inconsistent monotone longitudinal

reports of ever having sex. Analyses to assess the intervention effect on the age at first sex excluded the inconsistent records.

The percent utilizing the CAP varies more across the different calculation methods than for knowledge of the CAP. An interesting thing to note is that while the complete case % Util differs from the deterministically calculated percent by nearly 3%, (17.0% vs. 13.7%), the % Know does not change much (40.9% vs. 41.5%) and the three valid cell percents also differ by 1% or less. All MEMI models estimates for % Util are closer to the deterministic editing percent than they are to the complete case analysis estimate.

Editing Method	%Female	%US Born	%Had Sex	%Util	%Know	(Knowledge, Utilization) **A**(No, No)	**B**(Yes, No)	**D**(Yes, Yes)
Complete case	54.3	79.1	33.9	17.0	40.9	57.5	28.5	14.0
Deterministic	56.3	-	-	13.7	41.5	58.5	27.8	13.7
IRM / IML / IMV	56.3	82.0	34.8	14.5	42.2	56.2	28.0	15.8
Limited Edit	(55.0-57.5)	(81.0-83.0)	(34.3-35.3)	(14.0 - 15.0)	(41.5 - 42.9)	(55.5 - 56.6)	(27.3 - 28.3)	(15.3 - 16.1)
SyBRMICE	-	-	-	13.6	41.5	58.6	27.8	13.6
Ex 1: CAP	-	-	-	(13.1 - 14.1)	(40.8 - 42.1)	(57.9 - 59.3)	(27.2 - 28.5)	(13.1 - 14.1)
SyBRMICE	55.0	80.6	33.7	12.3	40.3	59.7	28.0	12.3
Ex 2: Full	(54.5 - 55.5)	(80.0 - 81.1)	(33.1 - 34.2)	(11.9 - 12.8)	(39.7 - 41.0)	(59.0 - 60.3)	(27.4 - 28.6)	(11.9 - 12.8)

Table 7.15: Estimated percents and 95% intervals for variables subject to editing and imputing calculated under different estimation methods.

7.4 Model Inference Comparison

This section assess the impact missing and inconsistent data can have on model inference. I fit a model that uses all the variables under consideration in this dissertation. I fit two models on the complete case (CC) data set, the Project Connect deterministically edited (DE) data set, and on each of the 20 MEMI data sets created from the SyBRMICE procedure. Model results from the MEMI data sets are combined and compared to the results from the CC and DE analyses.

The model was chosen to be similar to analyses performed by Project Connect researchers to assess the impact of the intervention on the likelihood a student would know about, or get a condom from the CAP. The models are not identical to any results published on this data, nor am I comparing my results directly to published results. Knowledge of the CAP is fit using the entire sample and utilization of the CAP is fit on the n_0 surveys from sexually experienced students only.

I fit two separate hierarchical logistic regression models (7.20), and (7.23) using `MCMCglmm` that includes a student level random effect to account for the repeated measures on some students. Predictor variables are ethnicity (AA, OTH), age (AGE), presence of a school-based health center (SBHC), intervention status (INTERV), gender (FEMALE), birthplace (US) and categorical study wave (T2, T3, T4). The model for knowledge also includes an indicator of ever having had sex (SEXP).

At this point I need to pause and explain how `MCMCglmm` handles logistic models. Hadfield (2010a) has structured the program to always include an over-dispersion parameter in generalized linear models

$$E[y] = \text{expit}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}).$$

When the data is Bernoulli, the value of the residual variance $\mathbf{e}$ is not defined and most other generalized linear mixed model packages in `R` fix $\mathbf{e}$ at 0, reducing the model to the standard logistic model. However, Hadfield (2010a) states that this is an arbitrary choice, and that the over dispersion parameter to allow the residual variance itself to vary, is the default. In addition `MCMCglmm` algorithm will not properly mix under this assumption (Hadfield, 2010b).

To work around this the residual variance was fixed at 1 as suggested by Hadfield (2010b,a).

The probability $p(y_{ij4} = 1) = \pi_{ij4}$ of a student reporting that they know they can get a condom from someone on campus is modeled using a logistic regression with predictors $\mathbf{X}_{ij}^{'}$ and a random intercept ζ_{i4} for each student.

$$\begin{aligned} y_{ij4}|\pi_{ij4} &\sim \text{Bernoulli}(\pi_{ij4}) \\ \text{logit}(\pi_{ij4}) &= \mathbf{X}_{ij}^{'}\boldsymbol{\alpha} + \zeta_{i4}, \end{aligned} \tag{7.20}$$

with priors

$$\begin{aligned} \zeta_{i4} &\sim \mathcal{N}(0, \sigma_4^2) \\ \sigma_4^2 &\sim IG(.001, .001), \end{aligned} \tag{7.21}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m_j$, and

$$\begin{aligned} \mathbf{X}_{ij} &= (1,\ \text{AA},\ \text{OTH},\ \text{AGE},\ \text{SBHC},\ \text{INTERV},\ \text{FEMALE},\ \text{US},\ \text{SEXP},\ \text{T2},\ \text{T3},\ \text{T4})^{'}), \\ \boldsymbol{\alpha} &\sim \mathcal{N}\left((-2, 0, 0, 0, 1, .5, 0, 0, .5, 0, .5, .5)^{'}, \frac{18415}{25}(X_k^{'}X_k)^{-1}\right), \end{aligned} \tag{7.22}$$

and where X_k is the covariate data matrix with rows $\mathbf{x_{ij}^{'}}$ for the model of knowledge from the complete cases. The model for utilization of the CAP is fit on the n_0 surveys where SEXP=1. For the complete cases $n_0 = 8,246$, across the 20 MEMI data sets n_0 ranges from 11,234 to 11,278.

$$\begin{aligned} y_{ij5}|\pi_{ij5} &\sim \text{Bernoulli}(\pi_{ij5}) \\ \text{logit}(\pi_{ij5}) &= \mathbf{W}_{ij}^{'}\boldsymbol{\beta} + \zeta_{i5}, \end{aligned} \tag{7.23}$$

with priors

$$\begin{aligned} \zeta_{i5} &\sim \mathcal{N}(0, \sigma_5^2) \\ \sigma_5^2 &\sim IG(.001, .001), \end{aligned} \tag{7.24}$$

for $i = 1, \ldots, n_0$ and $j = 1, \ldots, m_j$, and

$$\begin{aligned} \mathbf{W}_{ij} &= (1,\ \text{AA},\ \text{OTH},\ \text{AGE},\ \text{SBHC},\ \text{INTERV},\ \text{FEMALE},\ \text{US},\ \text{T2},\ \text{T3},\ \text{T4})^{'}), \\ \boldsymbol{\beta} &\sim \mathcal{N}\left((-3, 0, 0, 0, 2, .5, 0, 0, 0, .5, .5)^{'}, \frac{8246}{25}(W_u^{'}W_u)^{-1}\right), \end{aligned} \tag{7.25}$$

and where W_u is the covariate data matrix with rows $\mathbf{w_{ij}'}$ for the model of utilization from the complete cases of surveys on sexually experienced students. A weak proper prior is placed on the variance of the random effects for both models σ_4^2 and σ_5^2.

For the MEMI models the imputed and edited value for gender z_1 is used instead of the response value for FEMALE, similarly for US and SEXP. The prior means were specified to reflect the belief that there is a low probability of knowing about or utilizing the CAP, but having an SBHC, being in the intervention condition, having had sex, and during study years T3 and T4 increase this probability. The prior variances are calculated from the complete case data, but the same prior is used regardless of which data is being used in the model.

7.4.1 Results

Tables 7.16 and 7.17 display the sample size and frequency of predictors for the model on utilization and knowledge, respectively, under the complete cases, deterministically edited data set, and the MEMI data sets. Frequencies and percents that vary across MEMI data sets are presented in ***bold italics***. The model for knowledge is fit on all 23,796 surveys from high school students from the MEMI data sets, 18,415 (77.4%) complete cases, and 19,736 (82.9%) surveys from the deterministically edited data set. Since SEXP was multiply edited and imputed, the sample size for the MEMI data sets used to fit the model for Utilization ranges from 10,988 – 11,068 surveys, compared to 8,246 complete cases and 8,976 in the DE data set. Both models were run for 53,000 iterations with a burn-in period of 3,000 and a thin of 20, resulting in a final sample size of 2,500.

	Full High School Sample					
	Complete Case		Deterministic Edit		MEMI	
	N	%	N	%	N (Range)	% (Range)
Sample Size	18415	-	19736	-	23796	-
Sexually Experienced	8246	44.8%	8976	45.5%	***(10988 – 11068)***	***(46.2% – 46.5%)***
Knows about the CAP	7864	42.7%	8121	41.1%	***(9554 – 9644)***	***(40.1% – 40.5%)***
Got a condom from the CAP	2585	14.0%	3375	17.1%	***(2917 – 2954)***	***(12.3% – 12.4%)***
Female	10232	55.6%	10840	54.9%	***(12949 – 13016)***	***(54.4% – 54.7%)***
US Born	14400	78.2%	15415	78.1%	***(18715 – 18958)***	***(78.6% – 79.7%)***
Hispanic	14458	78.5%	15550	78.8%	18752	78.8%
African-American	1923	10.4%	2088	10.6%	2514	10.6%
Other	2034	11.0%	2098	10.6%	2530	10.6%
Age (Mean(SD))	16.14(1.30)	-	16.15(1.35)	-	16.19(1.29)	-
SBHC	9120	49.5%	9795	49.6%	11636	48.9%
Intervention	6297	34.2%	10409	52.7%	8332	35.0%
T1 (2005)	5032	27.3%	5242	26.6%	5922	24.9%
T2 (2006)	4696	25.5%	5079	25.7%	5831	24.5%
T3 (2007)	4395	23.9%	4765	24.1%	5879	24.7%
T4 (2008)	4292	23.3%	4650	23.6%	6164	25.9%

Table 7.16: Sample size and percents for predictors for the model of utilization under the complete cases, deterministically edited data set and the MEMI data sets. Frequencies and percents that vary across MEMI data sets are presented in ***bold italics***.

	Sexually Experienced Sample					
	Complete Case		Deterministic Edit		MEMI	
	N	%	N	%	N (Range)	% (Range)
Sample Size	8246	-	8976	-	***(10988 – 11068)***	-
Sexually Experienced						
Knows about the CAP	4114	49.9%	4252	47.4%	***(5022 – 5098)***	***(45.5% – 46.2%)***
Got a condom from the CAP	1897	23.0%	2407	26.8%	***(2100 – 2141)***	***(19.0% – 19.5%)***
Female	4146	50.3%	4453	49.6%	***(5422 – 5489)***	***(49.2% – 49.6%)***
US Born	6412	77.8%	6975	77.7%	***(8709 – 8862)***	***(79.0% – 80.3%)***
Hispanic	6452	78.2%	7051	78.6%	***(8640 – 8694)***	***(78.4% – 78.7%)***
African-American	1066	12.9%	1165	13.0%	***(1416 – 1439)***	***(12.8% – 13.0%)***
Other	728	8.8%	760	8.5%	***(923 – 953)***	***(8.4% – 8.6%)***
Age (Mean(SD))	16.57(1.18)	-	16.57(1.25)	-	***16.59(1.38)***	-
SBHC	4167	50.5%	4540	50.6%	***(5478 – 5520)***	***(49.7% – 50.0%)***
Intervention	2785	33.8%	4808	53.6%	***(3383 – 3916)***	***(35.2% – 35.5%)***
T1 (2005)	2349	28.5%	2464	27.5%	***(2749 – 2784)***	***(25.0% – 25.2%)***
T2 (2006)	2107	25.6%	2323	25.9%	***(2655 – 2685)***	***(24.1% – 24.3%)***
T3 (2007)	1932	23.4%	2145	23.9%	***(2713 – 2748)***	***(24.6% – 24.9%)***
T4 (2008)	1858	22.5%	2044	22.8%	***(2856 – 2889)***	***(25.9% – 26.2%)***

Table 7.17: Sample size and percents for predictors for the model of knowledge under the complete cases, deterministically edited data set and the MEMI data sets. Frequencies and percents that vary across MEMI data sets are presented in ***bold italics***.

Table 7.18 and 7.19 present the results for the regression models on knowledge and utilization of the CAP respectively. Estimates, standard errors and p-values are displayed for both complete case and MEMI analyses. For both models the standard errors are smaller for the MEMI results than for either the complete case or deterministic edited results. This is in part due to the increase in sample size. For nearly all coefficients the MEMI estimate is smaller than the CC or DE estimate, but mostly the inference about the parameters do not change. The cases where inference changes are discussed later.

Also presented in these tables are the number of standard deviations the MEMI estimate is from the CC and DE estimates. For example the the MEMI parameter estimate for the effect an SBHC has on a student's knowledge of the CAP was 5.9 CC standard deviations lower than the CC estimate, and 4.7 DE standard deviations lower than the DE estimate.

Most of the parameter estimates in the model of CAP knowledge differ between the two methods by over 1 standard deviation. However the indicators for T3 and T4 had a change in significance level. The percent of students in the complete case data set who knew about the CAP is 47.0% in T4 and 41.5% in T1, and this difference is significant ($p = .003$). Averaged across all MEMI data sets, these percents are 42.7% in T4 and 39.6% in T1, and are not significantly different ($p = .389$) after adjusting for other covariates. Similarly the percent of students in the deterministically edited data set who knew about the CAP is 44.0% in T3 and 40.4% in T1, and this difference is significant ($p < .001$). Averaged across all MEMI data sets, these percents are 43.4% in T4 and 39.6% in T1, and are not a significantly different ($p = .400$) after adjusting for other covariates.

Utilization parameter estimates of the intercept, presence of an SBHC, and study wave indicators in table 7.19 differ noticeably between the two models, however the only changes in significance interpretations occurred for the effects of wave (T2, T3 and T4). Waves T3 and T4 significantly differed from T1 when considering the observed, consistent cases. All three wave indicators differed from T1 when considering the deterministically edited data sample, but all three were not significantly different from T1 after multiply imputing and editing the missing and inconsistent data. Likely this is due to the increase in sample size of surveys on sexually experienced students for T3 and T4 compared to T1.

	Complete Cases			Deterministic Edit			Multiple Editing and Imputation using SyBRMICE			SD Diff	
Parameter	Estimate	SD	p	Estimate	SD	p	Estimate	SD	p	CC	DE
Intercept	−4.286	0.396	<0.001	−4.290	0.378	<0.001	−3.139	0.295	<0.001	2.9	2.9
African-American	−0.165	0.092	0.036	−0.196	0.091	0.020	−0.148	0.071	0.017	0.2	0.5
Other	0.399	0.088	<0.001	0.544	0.088	<0.001	0.400	0.069	<0.001	0.0	1.6
Age	0.096	0.023	<0.001	0.097	0.022	<0.001	0.056	0.017	<0.001	1.7	1.9
SBHC	2.686	0.093	<0.001	2.542	0.086	<0.001	2.139	0.061	<0.001	5.9	4.7
Intervention	0.655	0.068	<0.001	0.823	0.057	<0.001	0.515	0.051	<0.001	2.1	5.4
Female	−0.355	0.057	<0.000	−0.315	0.057	<0.001	−0.282	0.044	<0.001	1.3	0.6
US Born	0.527	0.072	<0.001	0.554	0.069	<0.001	0.433	0.062	<0.001	1.3	1.7
Has had sex	0.925	0.063	<0.001	0.796	0.058	<0.001	0.588	0.046	<0.001	5.3	3.6
T2	−0.589	0.082	<0.001	−0.482	0.074	<0.001	−0.486	0.062	<0.001	1.3	0.1
T3	0.128	0.082	0.065	−0.746	0.077	<0.001	−0.016	0.065	0.400	1.8	9.5
T4	0.225	0.083	0.003	−0.093	0.072	0.089	−0.017	0.063	0.389	2.9	1.1

Table 7.18: Results from a random intercept logistic regression model on CAP knowledge on the complete cases (CC), deterministic edited (DE) and MEMI data sets generated from the full SyBRMICE procedure. Differences that result in a change in significance are highlighted with a grey background. SD Diff is the number of CC or DE standard deviations away the MEMI estimate is from the CC or DE estimate.

	Complete Cases			Deterministic Edit			Multiple Editing and Imputation using SyBRMICE			SD Diff	
Parameter	Estimate	SD	p	Estimate	SD	p	Estimate	SD	p	CC	DE
Intercept	−5.103	0.699	<0.001	−3.289	0.667	<0.001	−3.726	0.598	<0.001	2.0	0.7
African-American	−0.288	0.150	0.023	−0.257	0.146	0.044	−0.247	0.124	0.022	0.3	0.1
Other	0.303	0.157	0.024	0.379	0.156	0.006	0.428	0.132	0.001	0.8	0.3
Age	0.054	0.038	0.076	−0.009	0.037	0.424	−0.012	0.033	0.362	1.7	0.1
SBHC	2.968	0.150	<0.001	2.803	0.147	<0.001	2.601	0.117	<0.001	2.4	1.4
Intervention	0.315	0.105	0.001	0.505	0.090	<0.001	0.285	0.090	0.001	0.3	2.5
Female	−0.325	0.088	<0.001	−0.375	0.089	<0.001	−0.298	0.078	<0.001	0.3	0.9
US Born	0.275	0.115	0.006	0.284	0.115	0.009	0.306	0.111	0.002	0.3	0.2
T2	0.135	0.130	0.147	−0.845	0.130	<0.001	0.005	0.113	0.515	1.0	6.5
T3	0.299	0.134	0.011	−0.343	0.128	0.001	−0.048	0.112	0.333	2.6	2.3
T4	0.502	0.133	<0.001	−0.260	0.125	0.022	−0.041	0.115	0.357	4.1	1.7

Table 7.19: Results from a random intercept logistic regression model on CAP utilization using the sexually experienced subset of the complete cases (CC), deterministic edited (DE) and MEMI data sets generated from the full SyBRMICE procedure. Differences that result in a change in significance are highlighted with a grey background. SD Diff is the number of CC or DE standard deviations away the MEMI estimate is from the CC or DE estimate.

7.5 Discussion

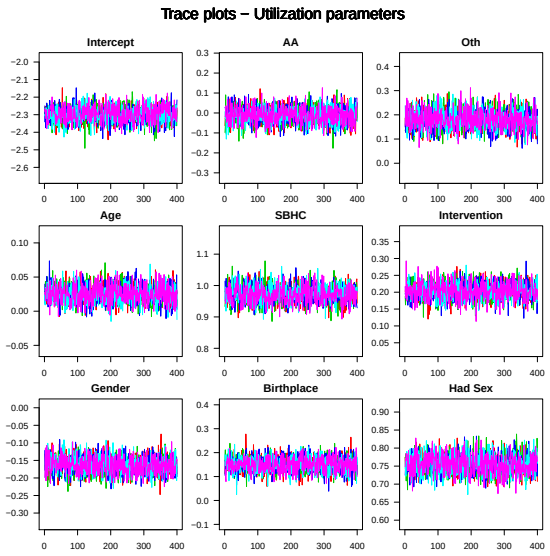
This chapter presents an all-inclusive example of how to take multiple variables subject to missing and/or inconsistent data and multiply edit and impute all data in a cyclical manner to produce multiple complete and consistent MEMI data sets. Results from subsequent analyses on the MEMI data sets can be combined to produce estimates and intervals that account for uncertainty in the imputation and editing processes.

The sample percentages in Table 7.15 clearly show the effect different imputation and editing procedures can have on these estimates. The model results comparison showed how compounding the effects of missing and inconsistent data not only in the outcome variable but also in the predictor variables can change model inferences. If there had been a lack of huge differences in this analysis it would not mean that the concept of multiply editing inconsistent data in a stochastic manner similar to multiple imputation shouldn't be done, just that the nuances and errors a data manager or analyst might notice did not create a large enough problem for this particular massive and complex data set.

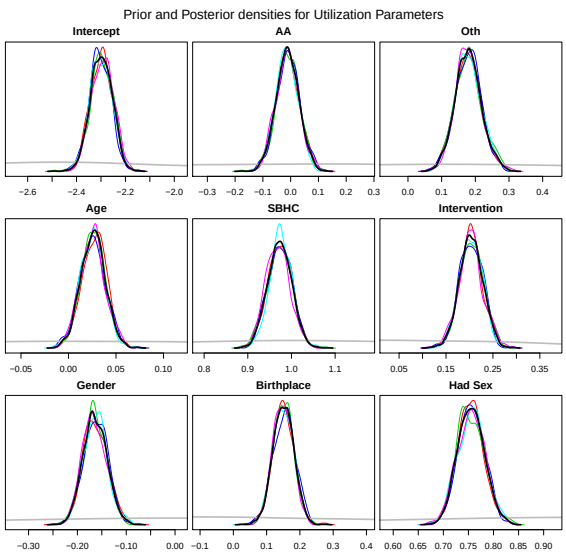
7.A Convergence Diagnostic Plots

Figures 7.3 – 7.4 display the trace, density, autocorrelation and Gelman-Rubin-Brooks diagnostic plots for the regression model parameters from the utilization and knowledge models respectively in the first example. Figures 7.5 – 7.9 display these same plots to assess model convergence for each of the 5 regression models in the second SyBRMICE example.

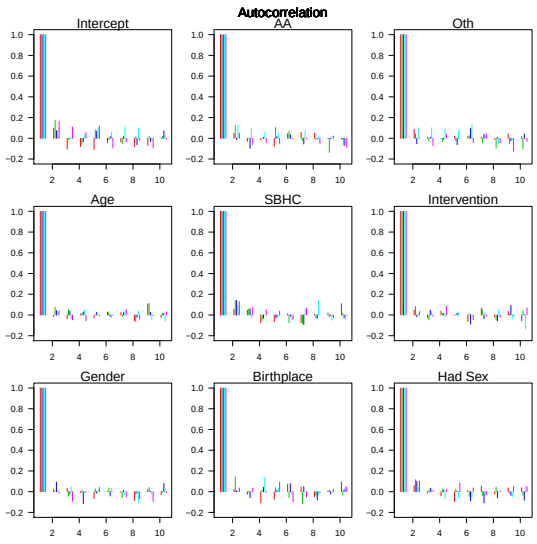
Plot (a) in each figure are the trace plots, which show that the chains are mixing well. Plot (b) plots the posterior densities of the model parameters, each one indicating that enough samples have been taken to create a smooth and approximately normally distributed posterior density. Plot (c) shows that the autocorrelation between subsequent samples after a thin of 20 is low enough to not be of concern, and plot (d) shows that the Gelman-Rubin-Brooks diagnostic plots indicate adequate convergence since the median value is around 1.



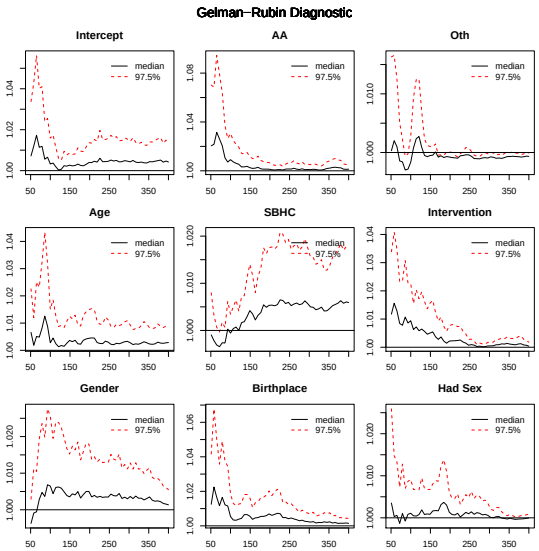
(a)



(b)

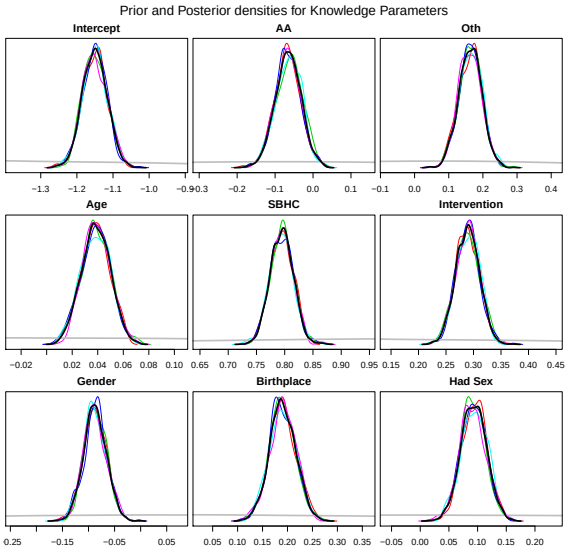


(c)

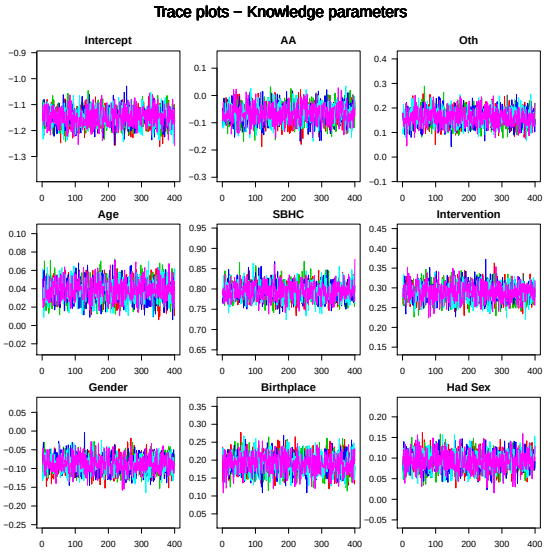


(d)

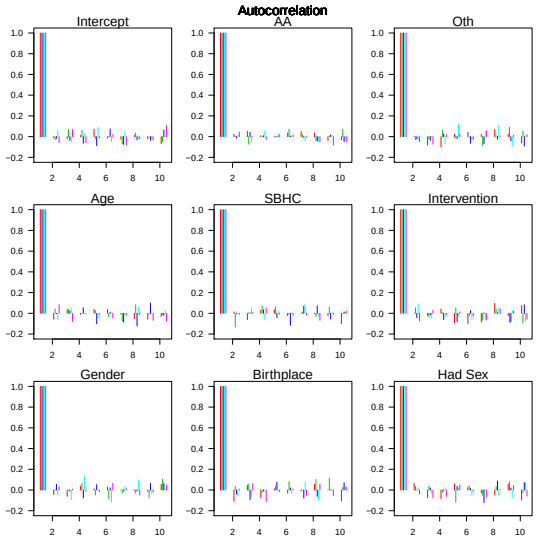
Figure 7.3: Trace and density plots for the utilization regression coefficients from example 1 in section 7.2.4. Prior densities are drawn in grey, each chain has its own color and the average density is drawn with a thick black line.



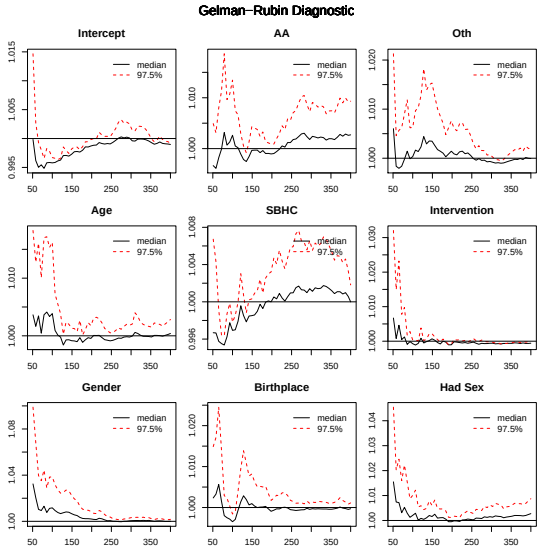
(a)



(b)

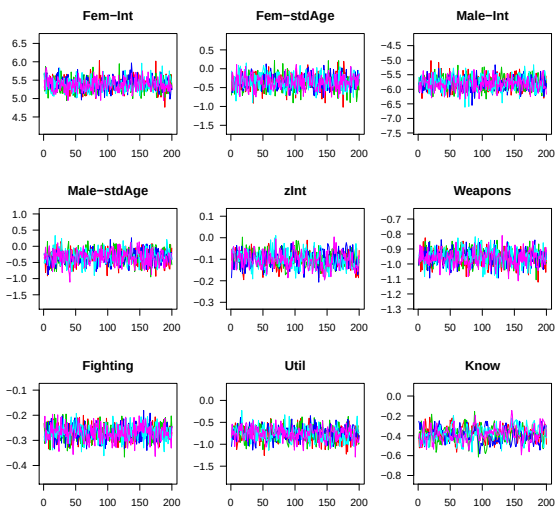


(c)

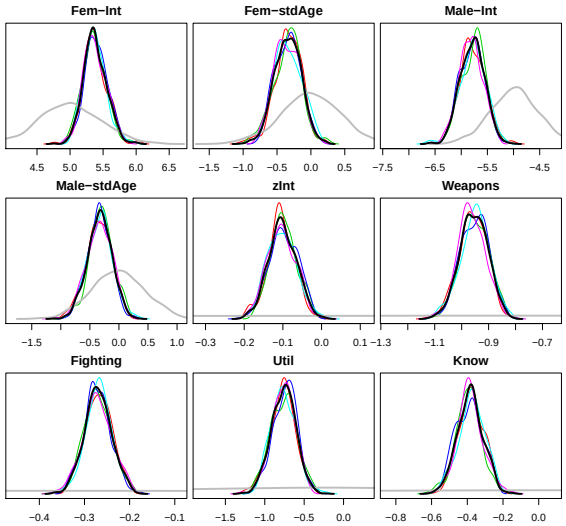


(d)

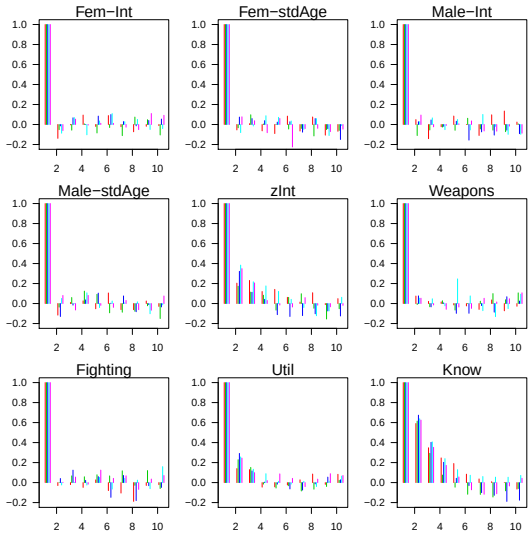
Figure 7.4: Trace and density plots for the knowledge regression coefficients from example 1 in section 7.2.4. Prior densities are drawn in grey, each chain has its own color and the average density is drawn with a thick black line.



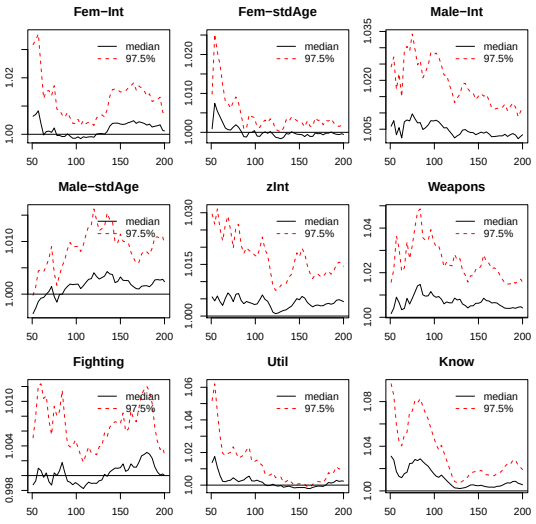
(a)



(b)

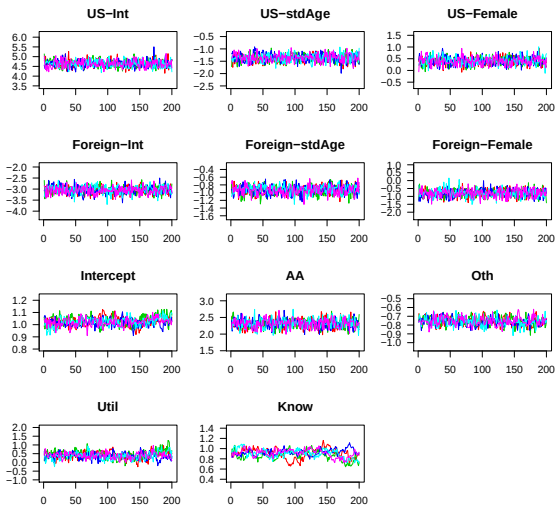


(c)

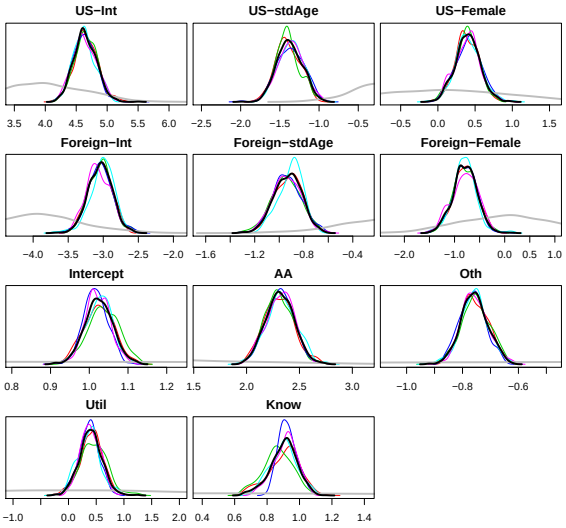


(d)

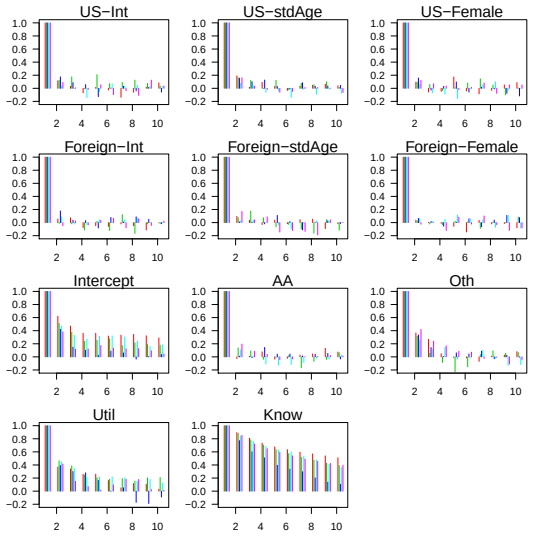
Figure 7.5: Convergence diagnostic plots for the IRM Gender regression parameters in the full SyBRMICE model.



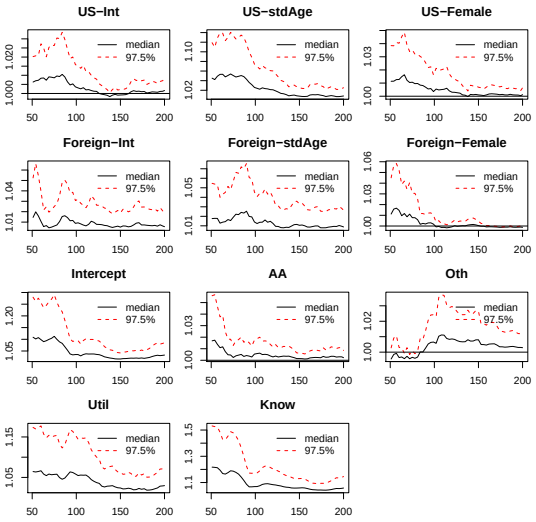
(a)



(b)

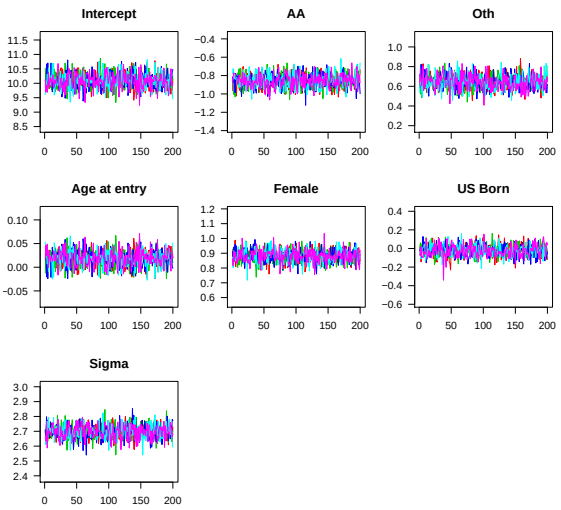


(c)

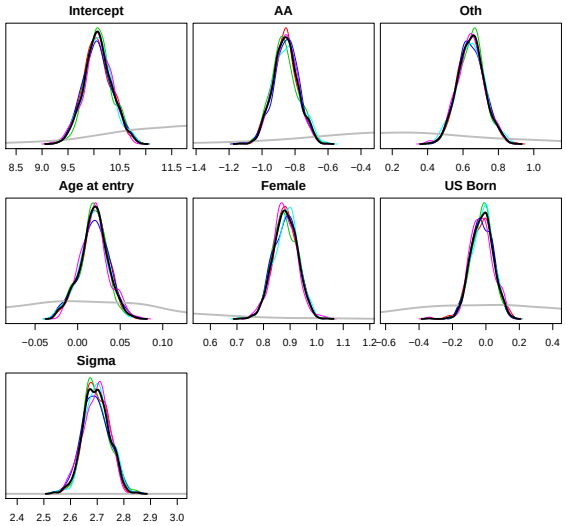


(d)

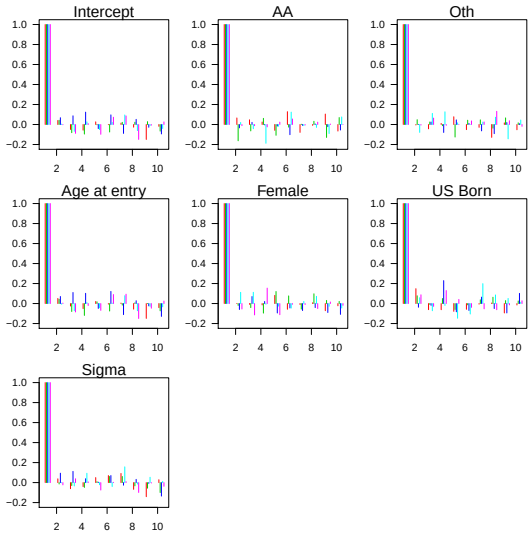
Figure 7.6: Convergence diagnostic plots for the IRM Birthplace regression parameters in the full SyBRMICE model.



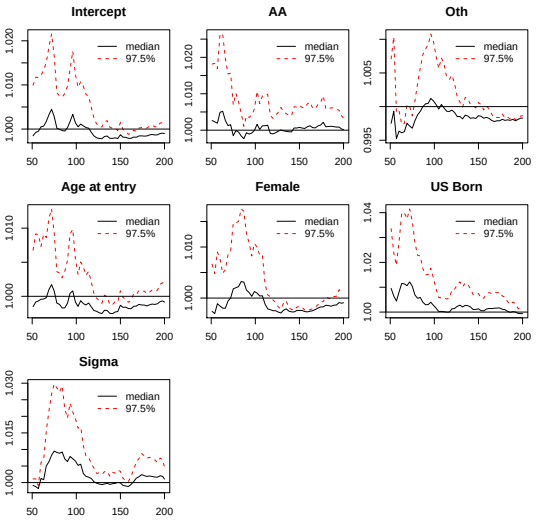
(a)



(b)

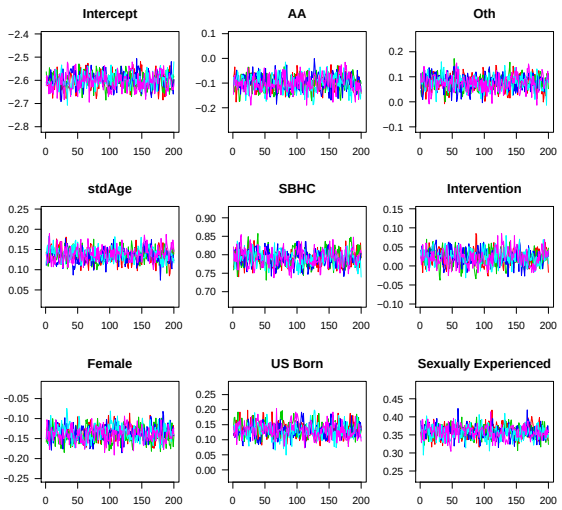


(c)

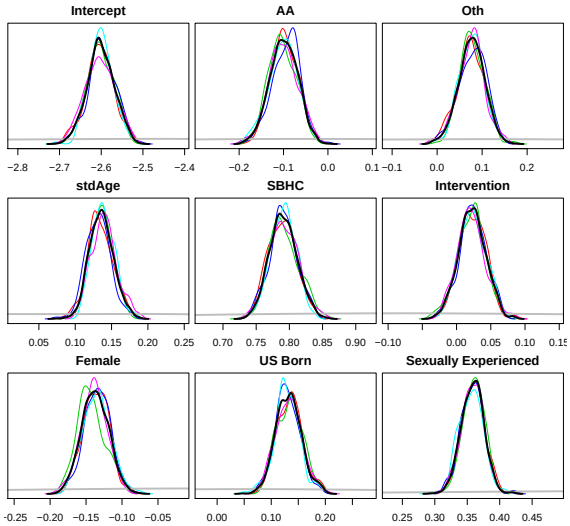


(d)

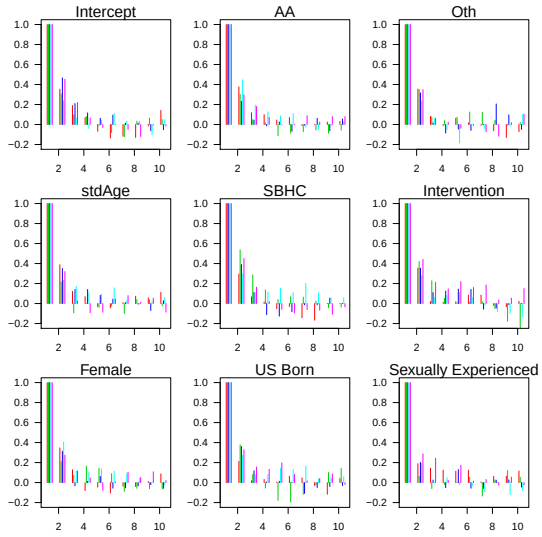
Figure 7.7: Convergence diagnostic plots for the IML grade at first sex regression parameters and variance Sigma in the full SyBRMICE model.



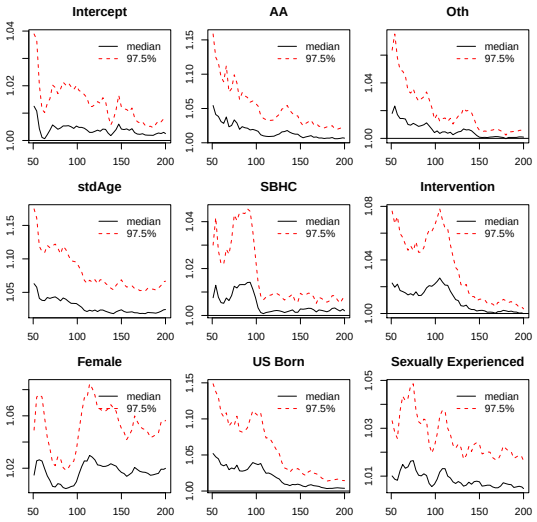
(a)



(b)

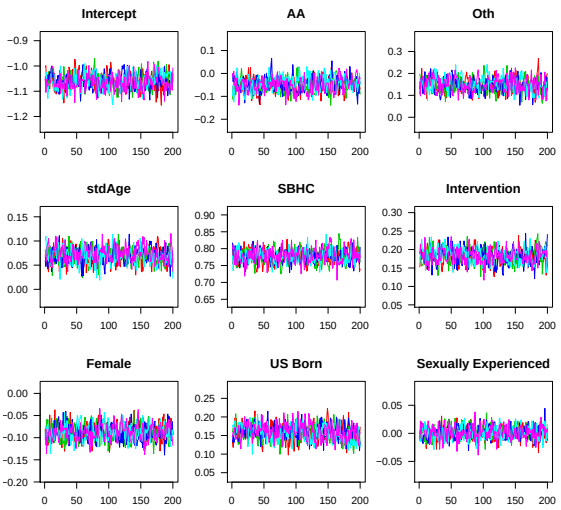


(c)

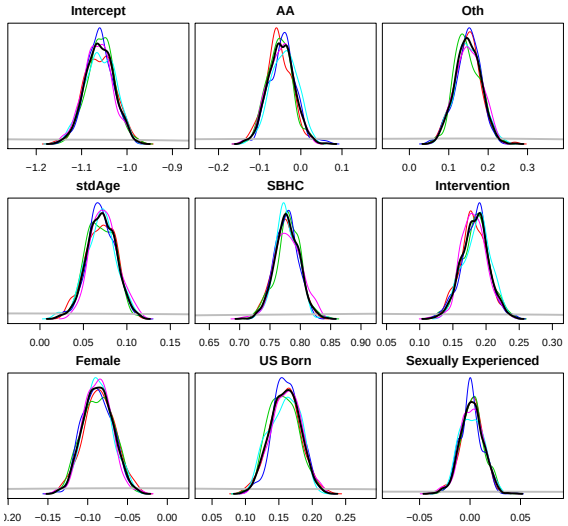


(d)

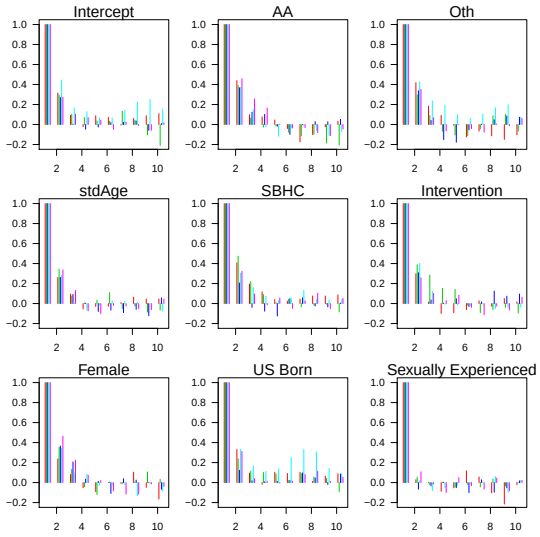
Figure 7.8: Convergence diagnostic plots for the IMV Utilization regression parameters in the full SyBRMICE model.



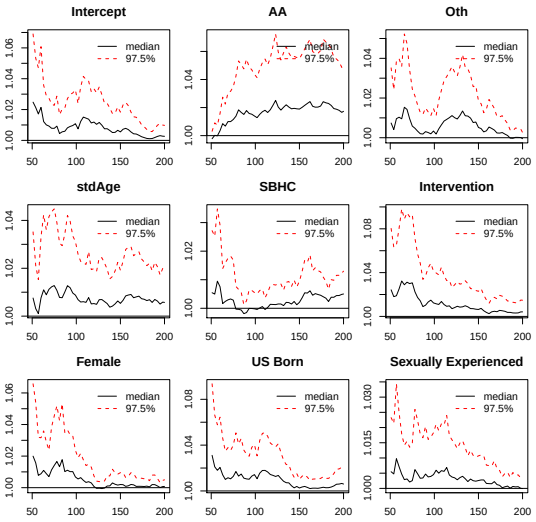
(a)



(b)



(c)



(d)

Figure 7.9: Convergence diagnostic plots for the IMV Knowledge regression parameters in the full SyBRMICE model.

CHAPTER 8

Conclusion

This dissertation introduced methods to multiply-edit inconsistent and erroneous data using methods similar to how missing data is treated under a multiple imputation framework. Uncertainty from the editing method is correctly propagated into the variance of the final results. The additional steps required to multiply-edit inconsistent data on top of a pre-existing multiple imputation procedure include needing to specify a model for the misspecification.

In Chapter 4 I demonstrated how to jointly impute and edit inconsistent repeated measures (IRM). These erroneous responses occur when the responses to the same question asked repeatedly over time differ across the multiple responses, when there really is only a single true response to that question that does not change over time. Examples were inconsistent reports of gender and of birthplace. The IRM model used a latent variable approach, but was seen to be computationally intensive. This likely is due more to an inefficient sampling routine than to structural problems in the model itself.

The model introduced in Chapter 5 looked at another type of inconsistent longitudinal response. The underlying true values either remain constant across time, or are allowed to change only once, and in a specific direction. I modeled the reports of ever having sexual intercourse, where consistent reporting patterns included always saying no, always saying yes, or saying no and then at some change point saying yes. The time to change point was modeled using an interval censored survival model and used to jointly impute and edit missing and inconsistent responses.

Chapter 6 introduced methods to impute and jointly edit inconsistent multivariate responses between two binary variables where a specific combination of responses gives mismatched or conflicting information. The variable combination I modeled was the reported

knowledge of and utilization of the Condom Availability Program on high school campuses and the inconsistent combination was a report of not knowing about the CAP but getting a condom from the program.

Lastly, in Chapter 7 I combined the repeated measures models, the longitudinal models and the multiple response models with the cyclical framework of SRMI (Raghunathan et al., 2001) to create SyBRMICE, a Sequential Bayesian Regression Model for Imputation and Conditional Editing. This example demonstrated that the SyBRMICE procedure extends easily to include any number of variables that may be subject to missing and/or inconsistent data. Furthermore, these multiple editing methods allow for the entire sample to be included in subsequent analyses.

Whether it be imputing missing data or editing inconsistent data, changing the data from its raw values will give you different results than what would be seen from the raw data. Especially with missing data, complete case analysis can give biased results when the missing data mechanism is not ignorable (Little and Rubin, 2002). Since inconsistent data occurs in so many forms, and the methods to change the incorrect data can be just as variable, there is no theoretical way to pre-determine if changing the data will give you significantly different estimates.

Some differences in the results from a regression model using the deterministically edited data and the multiply edited and multiply imputed data sets were seen here. Just as it is necessary to account for the error incurred by the imputation process, it is necessary to account for the error incurred from editing procedures.

Future work and extensions to the ideas presented in this dissertation could include the following. A prior distribution could be placed on the probability of making an error (π) in the IMV models of Chapter 6. A new method for the inconsistent monotone editing rules could be created to ensure minimal changes are made to the observed data rather than just sampling a new observation and completely ignoring the inconsistent data. Individuals with a lot of reporting errors and/or missing data could be flagged using some sort of criteria based on how variable their MEMI results are. This flag would indicate a need for further

examination to determine if they should be excluded from the data set entirely. The IMV framework could be used to model inconsistent monotone responses between two time points, and also extended to model more than two dichotomous variables.

Some may think that as we move more into the age of "Big Data", paper surveys and the lack of ability to enforce skip patterns and logical constraints will be a thing of the past. Survey deployment technology has come a long way from paper and pencil. There are a number of hand held devices specifically for field-survey use, but for the common researcher on a small to mid-sized grant these are may be unavailable. There are some free or relatively inexpensive methods to collect data on the web, but that requires a moderately tech-savvy person to correctly program and to administer. Nothing can, and likely ever will, beat the tried and true method of writing a survey in a word processing program and printing out copies for administration. This is especially the case in a school setting where you cannot depend on an internet connection, and there are heavy risks involved in bringing crates of electronic survey devices onto a school campus.

There will always be those who choose to falsify information to some extent regardless of how simple, easy and effective electronic surveys are to use. We can control electronics, but we can never control human nature. We can only edit and analyze their responses.

APPENDIX A

Codebook

Table A.1: Codebook for Section A: Demographics

Variable	Description	Values	Responses
A1.	Are you Male or Female?	0	Male
		1	Female
A2.	In what month were you born?	1	January
		⋮	⋮
		12	December
A3.	In what year were you born?	Middle School	
			1988-1995
		High School	
			1984-1991
A4.	How old are you?	Middle School	
			10-14
		High School	
			13-20

continued on next page

Variable	Description	Values	Responses
Mark all of the following that best describe you (A7)			
A71	African American/Black	0	Unmarked
		1	Marked
A72	Asian or Pacific Islander	0	Unmarked
		1	Marked
A73	Hispanic/Latino	0	Unmarked
		1	Marked
A74	Native American/American Indian	0	Unmarked
	/American Eskimo	1	Marked
A75	White/Caucasian	0	Unmarked
		1	Marked
A76	Other ethnicity	0	Unmarked
	(write-in allowed)	1	Marked
A9.	Where were you born?	1	United States
		2	Mexico
		3	El Salvador
		4	Guatemala
		5	China
		6	The Philippines
		7	Korea
	(write-in allowed)	8	Somewhere else

Table A.2: Codebook for Section I: Sexual Activity

Variable	Description	Values	Responses
I1.	Have you ever had sexual intercourse?	0	No
		1	Yes
I2.	How old were you when you had sexual intercourse for the first time?	Middle School	
		S	I have never had sex
		10	10 years old or younger
		⋮	⋮
		15	15 years old or older
		High School	
		S	I have never had sex
		10	10 years old or younger
		⋮	⋮
		17	17 years old or older
I3.	In what month did you have sexual intercourse for the first time?	S	I have never had sex
		1	January
		⋮	⋮
		12	December

**HS Only*

continued on next page

Variable	Description	Values	Responses
I4.	In what year did you have sexual intercourse for the first time?	Middle School	
		S	I have never had sex
			1999-2005
		High School	
		S	I have never had sex
			1994-2005
I5.	With how many people have you ever had sexual intercourse?	S	I have never had sex
		1	1 person
		2	2 people
		3	3 people
		4	4 people or more
I6.*	In the last 3 months, have you had sexual intercourse?	0	No
		1	Yes

**HS Only*

continued on next page

Variable	Description	Values	Responses
The last time you had sexual intercourse, did you or your partner use any of the following? Please check all that apply. (I7)			
I71	Condoms	0	Unmarked
		1	Marked
I72	Birth control pills	0	Unmarked
	patch or ring	1	Marked
I73	Birth control shots	0	Unmarked
		1	Marked
I74	Emergency contraception	0	Unmarked
	(morning after pill, plan B)	1	Marked
I75	Withdrawal (pull out)	0	Unmarked
		1	Marked
I76	Rhythm method	0	Unmarked
	(safe time of the month)	1	Marked
I77	Something else	0	Unmarked
		1	Marked
I78	Nothing	0	Unmarked
		1	Marked
I70	I have never had sexual intercourse	0	Unmarked
		1	Marked

**HS Only*

continued on next page

Variable	Description	Values	Responses
I8.*	Was the last person you had sexual intercourse with someone who you consider a steady or casual partner?		
		1	It was a steady partner
		2	It was a casual partner
		S	I have never had sex
I9.*	Counting all of the times you had sexual intercourse in the last 3 months, how often did you or your partner use a condom?		
		0	Never
		1	Less than half the time
		2	About half the time
		3	More than half the time
		4	Always
		S	I have not had sex in the last 3 months
I10.*	In the last 3 months, have you had sexual intercourse with someone that you have never had sexual intercourse with before?		
		0	No
		1	Yes

**HS Only* *continued on next page*

Variable	Description	Values	Responses
I11.	Have you ever given or received oral sex?	1	Yes, given only
		2	Yes, received only
		3	Yes, both
		0	No
		K	I don't know what oral sex is *(Added T2)*
I12.*	In the last 3 months, have you given or received oral sex?	1	Yes, given only
		2	Yes, received only
		3	Yes, both
		0	No
		K	I don't know what oral sex is *(Added T2)*
I13.*	Have you ever had anal sex?	0	No
		1	Yes
I14.*	In the last 3 months, have you had anal sex?	0	No
		1	Yes
I16.	With whom have you had any kind of sexual activity?	1	Males only
		2	Females only
		3	Males and females
		S	I have never had any
		S	kind of sexual activity

**HS Only*

Table A.3: Codebook for Section E: School Based Health Center

Variable	Description	Values	Responses
E1.*	Does your school have a health care clinic? This clinic also might be called the "teen clinic". It is different from the school nurse's office	0	No
		1	Yes
E2.*	Have you ever gone to the teen clinic at your school?	0	No
		1	Yes
What have you gone to the teen clinic at your school for? Please mark yes or no for EACH question (E3)			
E3a.*	Immunizations (shots)	0	No
		1	Yes
E3b.*	A check-up or sports physical	0	No
		1	Yes
E3c.*	Sickness (like a fever or infection)	0	No
		1	Yes
E3d.*	An injury (like a broken bone or cut)	0	No
		1	Yes

**HS Only*

continued on next page

Variable	Description	Values	Responses
E3e.*	Ongoing illness	0	No
	(like asthma or diabetes)	1	Yes
E3f.*	A check-up of my vagina or penis	0	No
		1	Yes
E3g.*	Birth control	0	No
		1	Yes
E3h.*	A test or treatment for a	0	No
	sexually transmitted disease (STD)	1	Yes
E3i.*	Counseling	0	No
		1	Yes
E3j.*	Information about sex	0	No
		1	Yes
E3k.*	Information about my health	0	No
		1	Yes
E3l.*	Something else	0	No
		1	Yes

**HS Only*

continued on next page

Variable	Description	Values	Responses
What has kept you from using the teen clinic at your school? Please check all that apply. (E4)			
E41.*	I didn't feel comfortable	0	Unmarked
		1	Marked
E42.*	My parents didn't give permission	0	Unmarked
		1	Marked
E43.*	I didnt know where it was or how to make an appointment	0	Unmarked
		1	Marked
E44.*	I thought Id have to pay	0	Unmarked
		1	Marked
E45.*	I wouldn't want anyone to know I went there	0	Unmarked
		1	Marked
E46.*	I don't have any health problems	0	Unmarked
		1	Marked
E47.*	I have my own doctor	0	Unmarked
		1	Marked
E48.*	Nothing has kept me from using the teen clinic	0	Unmarked
		1	Marked

**HS Only*

Table A.4: Codebook for Section G: Condom Availability Program

Variable	Description	Values	Responses
G1	Does someone at your school (like the school nurse or a counselor)		
	give out condoms to students	0	No
	who want them?	1	Yes
		K	Don't know
	Skip pattern text on survey removed after T1		
Do any of the following people give out condoms at your school?			
Please mark yes or no for EACH question(G2)			
G2a.	Health clinic staff or School Nurse	0	No
		1	Yes
	(Option added in T2)	K	Don't know
G2b.	An administrator	0	No
		1	Yes
	(Option added in T2)	K	Don't know
G2c.	Health Teacher	0	No
		1	Yes
	(Option added in T2)	K	Don't know
G2d.	A PE Teacher or Coach	0	No
		1	Yes
	(Option added in T2)	K	Don't know

Continued on next page

Variable	Description	Values	Responses
Who gives out condoms?(G2) cont.			
G2e.	Someone else	0	No
		1	Yes
	(Option added in T2)	K	Don't know
G2f.	No one gives out condoms at my school*(Question added T2)*	0	No
		1	Yes
		K	Don't know
G3.	Have you ever gotten condoms from this person at your school?	0	No
		1	Yes
G4.	In the past month, how many times have you gotten condoms from this person at your school?		
		0	Never
		1	One
		2	Two
		3	Three or more

A.1 Sexual Activity Recode Rules

Table A.5: Deterministic editing rules currently applied to the sexual activity section of the Project Connect survey data.

#	**Recode Rule**
1	If I1 is M and I2 through I5 are all S then change I1 to No
2	If I1 is M and I2-I5 all M with at least one S then change I1 to No and I2-I5 to S
3	If I1-I5 all M, and I16 is S then change I1 to No and I2-I5 to S
4	If I1 is M and I2-I5 consist of a combination of non-missing and M then change I1 to Yes
5	If I1 is M and I2-I5 consist of a combination of non-missing and S then change I1-I5 to M
6	If I1 is No and I2-I5 consist of a combination of non-missing and S then change I1-I5 to M
7	If I1 is No and at least 3 of I2-I5 are non-missing with the others missing then change I1 to Yes
8	If I1 is No and I2-I5 all M then change I2-I5 to S
9	If I1 is No and I2-I5 all S or M with only one non-missing then change the non-missing answer to S
10	If I1 is No and I2-I5 are all combinations of S and M then change I2-I5 to S
11	If I1 is Yes and any of I2-I5 is S then change I1-I5 to M
12	If I1 is Yes and I16 is S then change I16 to M
13	If I1 is Yes and I8 is S then change I8 to M
14	If I1 is Yes and I70 is marked then change I70 to Unmarked
15	If more than 5 of I71-I78 are marked then change I70 - I78 to M
16	If I72, I73 & I74 are all marked then change I70 - I78 to M

M = Missing, S = I have not had sex, K=Don't Know, I = Inconsistent

Table A.5: Deterministic editing rules currently applied to the sexual activity section of the Project Connect survey data.

#	**Recode Rule**
17	If I72 and I73 are both marked then change I70 - I78 to M
18	If I11 is M and I12 is Yes or K then change I11 to the matching Yes answer
19	If I11 is No and I12 is any of the Yes answers then change both to I
20	If one of I11 or I12 is K and the other one is a non-missing answer then change both to I
21	If I11 is given only and I12 is received only or both then change both to I
22	If I11 is received only and I12 is given only or both then change both to I
23	If I13 is M and I14 is Yes then change I13 to Yes
24	If I13 is No and I14 is Yes then change both to I
25	If I6 is Yes and I9 is S or if I6 is No and I9 has a non-missing answer then change both to I
26	If I2 is greater than the reported age then change I2 to M

M = Missing, S = I have not had sex, K=Don't Know, I = Inconsistent

Bibliography

Albert, J. (2009), *Bayesian Computation with R*, Springer.

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

Brener, N. D., Collins, J. L., Kann, L., Warren, C. W., and Williams, B. I. (1995), "Reliability of the Youth Risk Behavior Survey Questionnaire," *American Journal Epidemiology*, 141, 575–580.

Brooks, S. and Gelman, A. (1998), "General Methods for Monitoring Convergence of Iterative Simulations," *Journal of Computational and Graphical Statistics*, 7, 434–455.

Bruni, R., Reale, A., and Torelli, R. (2002), "DIESIS: a New Software System for Editing and Imputation," in *Proceedings of 41st Riunione Scientifica SIS 2002*.

Cannell, C., Miller, P., and Oksenberg, L. (1981), "Research on Interviewing Techniques," in *Sociological Methodology*, ed. S. Leinhard, San Francisco, CA: Jossey-Bass, pp. 389–437.

Centers for Disease Control and Prevention (2010), "Youth Risk Behavior Surveillance Summaries," *MMWR*, 59, SS–5.

Charlton, J. (2003), "Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit Project, Volume 1," `http://www.cs.york.ac.uk/euredit/results/results.html/`.

Christensen, R., Johnson, W., and Branscum, A. (2010), *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, Chapman & Hall/CRC.

Cole, S. R., Chu, H., and Greenland, S. (2006), "Multiple Imputation for Measurement-Error Correction," *International Journal of Epidemiology*, 35, 1074–1081.

Daniels, M. J. and Hogan, J. W. (2007), *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*, Chapman & Hall/CRC.

Davey, A., Shanahan, M. J., and Schafer, J. L. (2001), "Correcting for Selective Nonresponse in the National Longitudinal Survey of Youth using Multiple Imputation," *The Journal of Human Resources*, 36, 500–519.

de Waal, T. and Coutinho, W. (2005), "Editing for Business Surveys: An Assessment of Selected Algorithms," *International Statistical Review*, 73, 73–102.

Demirtas, H. and Hedeker, D. (2007), "Gaussianization-based Quasi-imputation and Expansion Strategies for Incomplete Correlated Binary Responses," *Statistics in Medicine*, 26, 782–799.

— (2008), "An Imputation Strategy for Incomplete Longitudinal Ordinal Data," *Statistics in Medicine*, 27, 4086–4093.

DeRosa, C. J., Jeffries, R. A., Afifi, A. A., Cumberland, W. G., Chung, E. Q., Kerndt, P. R., Ethier, K. A., Martinez, E., Loya, R. V., and Dittus, P. J. (2012), "Improving the Implementation of a Condom Availability Policy in Urban High Schools: Impact on Student Awareness and Utilization of the Program," *Journal of Adolescent Health*, 51, 572–579.

Ethier, K. A., Dittus, P. J., DeRosa, C. J., Chung, E. Q., Martinez, E., and Kerndt, P. R. (2011), "School-Based Health Center Access, Reproductive Health Care, and Contraceptive Use Among Sexually Experienced High School Students," *Journal of Adolescent Health*, 48, 562–565.

Fellegi, I. P. and Holt, D. (1976), "Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association*, 71, 17–35.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association*, 85, 972–985.

Gelfand, A. E. and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

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

Geman, S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Ghosh-Dastidar, B. and Schafer, J. L. (2003), "Multiple Edit/Multiple Imputation for Multivariate Continuous Data," *Journal of the American Statistical Association*, 98, 807–817.

Goodkind, T. (2001), *Faith of the Fallen*, Sword of Truth, Tom Doherty Associates.

Habel, M. A., Dittus, P. J., DeRosa, C. J., Chung, E. Q., and Kerndt, P. R. (2010), "Sports Participation and Student's Sexual Activity," *Perspectives on Sexual and Reproductive Health*, 42, 244–250.

Hadfield, J. (2010a), *MCMC Generalised Linear Mixed Models - Course Notes*, [Online; accessed May-2010].

— (2010b), "MCMCglmm help - information about 'units' term," `https://stat.ethz.ch/pipermail/r-sig-mixed-models/2010q3/004006.html/`, [Online; R-SIG-ME Listserv, accessed 16-Nov-2012].

Hadfield, J. D. (2010c), "MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package," *Journal of Statistical Software*, 33, 1–22.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and their Applications," *Biometrika*, 57, 97–109.

Herzog, T., Scheuren, F., and Winkler, W. (2007), *Automatic Editing and Imputation of Sample Survey Data*, Springer.

Horton, N. J. and Kleinman, K. P. (2007), "Much Ado about Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models," *The American Statistician*, 61, 79–90.

Knaus, J. (2010), *snowfall: Easier Cluster Computing (based on snow)*, R package version 1.84.

Kovar, J. G., MacMillan, J. H., and Whitridge, P. (1991), "Overview and Strategy for the Generalized Edit and Imputation System," in *Statistics Canada, Methodology Branch Working Paper BSMD 88-007E.*

Link, W. A. and Eaton, M. J. (2012), "On Thinning of Chains in MCMC," *Methods in Ecology and Evolution*, 3, 112–115.

Little, R. J. A. (1988), "A Test of Missing Completely at Random for Multivariate Data with Missing Values," *Journal of the American Statistical Association*, 83, 1198–1202.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1092.

Müller, P. (1991), "A Generic Approach to Posterior Integration and Gibbs Sampling," Tech. Rep. 91-09, Purdue University.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rafferty, Y. and Radosh, A. (1997), "Attitudes about AIDS Education and Condom Availability among Parents of High School Students in New York City: A Focus Group Approach." *AIDS Education and Prevention*, 9, 14–30.

Raghunathan, T. E. (2004), "What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data," *Annual Review of Public Health*, 25, 99–117.

Raghunathan, T. E., Lepkowski, J. M., and Hoewyk, J. V. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," *Survey Methodology*, 27, 85–95.

Rancourt, E. (2001), "Edit and Imputation: From Suspicious to Scientific Techniques," in *53rd session of the International Statistical Institute, Seoul, Korea*.

Reiter, J. P. and Raghunathan, T. E. (2007), "The Multiple Adaptations of Multiple Imputation," *Journal of the American Statistical Association*, 102, 1462–1470.

Robert, C. P. and Casella, G. (2005), *Monte Carlo Statistical Methods*, Springer.

Royston, P. (2004), "Multiple Imputation of Missing Values," *Stata Journal*, 4, (15)227–241.

Royston, P. and White, I. R. (2011), "Multiple Imputation by Chained Equations (MICE): Implementation in Stata," *Journal of Statistical Software*, 45, 1–20.

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

— (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley.

Rubin, D. B. and Schafer, J. L. (1990), "Efficiently Creating Multiple Imputations for Incomplete Multivariate Normal Data," in *Proceedings of the Statistical Computing Section of the American Statistical Association*, pp. 83–88.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall/CRC.

Tang, L., Song, J., Belin, T. R., and Unützer, J. (2005), "A Comparison of Imputation Methods in a Longitudinal Randomized Clinical Trial," *Statistics in Medicine*, 24, 2111–2128.

Tierney, L., Rossini, A. J., Li, N., and Sevcikova, H. (2011), *snow: Simple Network of Workstations*, R package version 0.3-8.

Troped, P. J., Wiecha, J. L., Fragala, M. S., Matthews, C. E., Finkelstein, D. M., Kim, J., and Peterson, K. E. (2007), "Reliability and Validity of YRBS Physical Activity Items Among Middle School Students," *Medicine and Science in Sports and Exercise*, 39, 416–425.

United Nations Economic Commision for Europe (2011), "Conference of European Statisticians: Work Session on Statistical Data Editing," in *Conference of European Statisticians: Work Session on Statistical Data Editing*.

van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999), "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis," *Statistics in Medicine*, 18, 681–694.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011), "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, 45, 1–67.

White, I. (2006), "Commentary: Dealing With Measurement Error: Multiple Imputation or Regression Calibration?" *International Journal of Epidemiology*, 35, 1081–1082.

Winkler, W. E. (2003), "A Contingency-Table Model for Imputing Data Satisfying Analytic Constraints," Statistical Research Division Research Report RRS2003/07, US Census Bureau, Washington D.C.

Winkler, W. E. and Draper, L. R. (1996), "Application of the SPEER Edit System," Statistical Research Division Research Report RR96/02, US Census Bureau, Washington D.C.

Winkler, W. E. and Petkunas, T. F. (1997), "The DISCRETE edit system," in *Proceedings of the Conference of European Statisticians, Section on Statistical Data Editing, UNECE*, pp. 51–55.

Yang, X., Li, J., and Shoptaw, S. (2008), "Imputation-Based Strategies for Clinical Trial Longitudinal Data with Nonignorable Missing Values," *Statistics in Medicine*, 27, 2826–2849.

Yang, X. and Shoptaw, S. (2005), "Assessing Missing Data Assumptions in Longitudinal Studies: An Example Using a Smoking Cessation Trial." *Drug and Alcohol Dependence*, 77, 213–225.

Yucel, R. M. (2011), "State of the Multiple Imputation Software," *Journal of Statistical Software*, 45, 1–7.

Zellner, A. (1983), "Applications of Bayesian Analysis in Econometrics," *Journal of the Royal Statistical Society, Series D (The Statistician)*, 32, 23–34.