Analysis of Longitudinal Surveys with Missing Responses

by

Iván Adolfo Carrillo García

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2008

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Longitudinal surveys have emerged in recent years as an important data collection tool for population studies where the primary interest is to examine population changes over time at the individual level. The National Longitudinal Survey of Children and Youth (NLSCY), a large scale survey with a complex sampling design and conducted by Statistics Canada, follows a large group of children and youth over time and collects measurement on various indicators related to their educational, behavioral and psychological development. One of the major objectives of the study is to explore how such development is related to or affected by familial, environmental and economical factors.

The generalized estimating equation approach, sometimes better known as the GEE method, is the most popular statistical inference tool for longitudinal studies. The vast majority of existing literature on the GEE method, however, uses the method for non-survey settings; and issues related to complex sampling designs are ignored.

This thesis develops methods for the analysis of longitudinal surveys when the response variable contains missing values. Our methods are built within the GEE framework, with a major focus on using the GEE method when missing responses are handled through hot-deck imputation. We first argue why, and further show how, the survey weights can be incorporated into the so-called Pseudo GEE method under a joint randomization framework. The consistency of the resulting Pseudo GEE estimators with complete responses is established under the proposed framework.

The main focus of this research is to extend the proposed pseudo GEE method to cover cases where the missing responses are imputed through the hot-deck method. Both weighted and unweighted hot-deck imputation procedures are considered. The consistency of the pseudo GEE estimators under imputation for missing responses is established for both procedures. Linearization variance estimators are developed for the pseudo GEE estimators under the assumption that the finite population sampling fraction is small or negligible, a scenario often held for large scale population surveys.

Finite sample performances of the proposed estimators are investigated through an extensive simulation study. The results show that the pseudo GEE estimators and the linearization variance estimators perform well under several sampling designs and for both continuous response and binary response.

KEYWORDS: Longitudinal surveys, Complex surveys, GEE, Pseudo-GEE, Missing values, Weighted GEE, Hot-deck imputation, Consistency of Pseudo-GEE estimators (with hot-deck imputation), Variance estimation, Joint randomization.

Acknowledgements

I would like to take this opportunity to thank my supervisor, Changbao Wu, for his excellent mentoring throughout my PhD studies. I am grateful for his encouragement and great patience especially when my progress was at a slow pace; which is, most of the time. Without his advising and continued help this thesis would not be. Also, his contagious passion for research and teaching has been and will always inspirit me to look for excellence in my career. I would like to thank him for the financial support during my studies, too.

I thank my co-supervisor, Jiahua Chen, for the many good suggestions to improve this thesis. Thanks to them this thesis is much easier to follow and read. I also appreciate him walking me through some of the hardest parts of the proofs and developments so that I could gain better understanding of what many times I just followed almost mechanically.

The questions, comments, and suggestions from the professors in my committee, Mary Thompson and Grace Yi, have been really helpful for enlarging the usefulness of this research; I thank them very much. Additionally, I want to thank Jock MacKay for his willingness to be Grace's delegate in my internal defence; he thereupon suggested some important points in an earlier version of this thesis.

I would also like to thank Dr. Milorad Kovacevic, who was a great mentor during my internship at Statistics Canada. A lot of what he helped me with is included in this thesis. I thank Statistics Canada, MITACS, and the NPCDS for making my internship at Statistics Canada possible.

And lastly, I am grateful to the free software community; this thesis was almost entirely developed using free (as in free speech) software. I am particularly indebted to the developers, maintainers, and contributors of GNU/Linux, LaTeX, Kile, Gnumeric, Emacs, R, and RKWard.

Contents

List of Tables

Chapter 1

Introduction

1.1 Longitudinal Studies

There exist two major types of statistical research designs, namely, cross-sectional studies and longitudinal studies. Cross-sectional studies can be described as "one-time" or "one-shot" studies. Here, interest lies in the characteristics of a certain population or model at a particular time point; subjects and variables of interest are considered only in reference to that time. On the other hand, in longitudinal studies, also called "panel studies", the investigator is interested in some aspect of a population through time. In this case, the variables of interest are measured on a fixed set of units at several time points during the reference time period.

According to Kish (1987), the major advantage of longitudinal studies over cross-sectional is that they allow for measuring gross, or micro, changes for units in a population; it is possible to estimate the distribution of individual changes. Whereas longitudinal studies are designed to go beyond the measuring of current levels or net (macro) changes. Kish goes on to argue that without a longitudinal study, some gross changes can be masked behind net changes; and that "averages and sums of repeated samples can lead to better statistical inference than a 'one-shot' sample."

Diggle et al. (2002) and Hedeker and Gibbons (2006) point out that, with longitudinal studies, contrary to a cross-sectional study, it is possible to separate age and cohort effects. Where age effect is the actual change within subjects over time, and cohort effect is the difference between units at the beginning of the study period.

Hedeker and Gibbons (2006) also suggest that since longitudinal studies allow for the measurement of time-varying explanatory variables (covariates), the statistical inferences about dynamic relationship between the outcome on interest (response) and these covariates are much stronger than those based on cross-sectional studies.

When we are interested in the marginal mean of a variable, possibly conditionally on some covariates, and not in measuring change, a longitudinal study is not necessary; a cross-sectional study suffices. However, even in this case, a longitudinal study tends to be more powerful, because each subject serves as his or her own control for any unmeasured characteristics (Diggle et al., 2002).

Some longitudinal studies can be avoided by using a "retrospective" design instead. In this

case, a cross-sectional set of units is selected, and some variables are "measured" backwards in time with the help of the subject's memory or records. Nonetheless, Korn and Graubard (1999) indicate that some subjects may not recall their past information accurately; and that with a longitudinal study there is less recall error. They also note that some variables of interest may require actual measurements (like blood chemistries) that are usually not available from the past.

The advantages of longitudinal studies, however, do not come without a price. On the one hand, there are the operational constraints, such as higher costs and longer completion time than for a cross-sectional study (Korn and Graubard, 1999). And on the other hand, there is some added complexity in analyses, like the need to take into consideration the lack of independence among responses coming from the same unit; and also some data quality difficulties.

One of the problems with longitudinal studies is that the composition of many populations changes over time (Duncan and Kalton, 1987). Another disturbing factor is that the measurement instrument(s) may change over time (Kish, 1987). These authors also discuss other data quality issues in longitudinal studies, like panel conditioning, panel bias, panel contamination, sensitizing, and learning.

Another big problem with longitudinal studies is missing values. This problem is particularly common in surveys, even for cross-sectional ones (see, for example Groves et al., 2002), and the problem intensifies for longitudinal surveys. As Song (2007) puts it, "it is more difficult to deal with missing data in longitudinal studies. This is because missing data patterns appear much more sophisticated than those in cross-sectional studies." The usual kinds of missing values in cross-sectional studies are unit nonresponse, missing variable of interest, and missing covariate(s). Longitudinal studies suffer these too; but *in addition, also* contain other types of missing values: attrition (or drop-outs), intermittent missingness, and any combination of all the types.

The next section presents a large scale survey conducted by Statistics Canada, the NLSCY. This survey motivates in part this thesis and will be used for analysis and simulations.

1.2 The National Longitudinal Survey of Children and Youth (NLSCY)

The National Longitudinal Survey of Children and Youth (NLSCY) is a longitudinal survey by Human Resources Development Canada designed to measure child development and well-being. The main objective of the survey is to study the development of children's behaviour problems as they grow as well as examining the factors that contribute to change. It consists of (so far) six biennial cycles conducted from 1994 to 2005, and looked at households with children from 0 to 11 years old at the first cycle. An in depth description of the NLSCY features is given in Appendix A.

One very important measurement of the NLSCY is the aggressive behaviour of young children. "Aggression in childhood has been linked with later aggression, delinquency, and crime in adolescence and adulthood; with poor school outcomes; with unemployment in adulthood;

and with other negative circumstances" (Thomas, 2004). This will be the outcome of interest in some of our simulations. The response variable is the "Physical Aggression Score," (PAS). This variable is a scale from 0-12, based on eight or six questions (depending on the age); a high score indicates behaviours associated with conduct disorders, physical aggression, and opposition. PAS is scaled from 0 to 16 based on eight questions for children who are 2 to 3 years old, and is scaled from 0 to 12 based on six questions for children who are 4 to 11 years old. For the results to be comparable across different age groups, PAS's are unified to a scale of 0 to 12. To do this we simply multiply the score for 2-3 year-old children by 12/16=0.75 and leave the score for 4-11 year-old children unchanged. Although PAS is an ordinal variable, it is reasonable to treat it as continuous since it has more than seven categories (Carrillo et al., 2005); we do so in some of our simulations, although in other simulations we categorize it to only two levels (for logistic regression).

Given that this question is asked only for kids who are 2 to 11 years old, and we would like to apply the methods to a *longitudinal* sample, we will restrict our analyses to the first four cycles, even though there are six cycles of data available. This is so because by the fifth cycle (i.e. approximately 8 years after cycle 1) most of the longitudinal kids surveyed at cycle 1 (who were asked that question) will be out of scope for the PAS. Additionally, we will restrict to kids who were 2 to 5 years old at cycle 1 because most of those kids who were 6 or older are out of scope by cycle 4. There are 7,637 kids who responded in cycle 1 with these restrictions (2-5). However not all of them were selected to be in the longitudinal sample; only 5,610 kids are "longitudinal" in cycle 1. Two of these kids, who were 2 years old at cycle 1, appear as being 2 or 3 years old at cycle 2. As this is really strange since the surveys are two years apart, and those are the only two kids who were less than 4 years old at cycle 2, we decided to exclude them in order to avoid issues of outliers. In other words we want to avoid these two kids unduly influencing the estimators since these observations are likely to be different from the rest of kids in the cohort. Additionally, 38 kids who were 5 years old at cycle 1 became 12 years old by cycle 4, and thus out of scope; we also excluded these kids. So, we are left with 5,570 kids who were 2-5 years old at cycle 1 and 9-11 years old at cycle 4. So we need to redefine our population of interest as those children who were 2-5 years old from 1994 to 1995 *and* 9-11 years old from 2000 to 2001.

Table 1.1 shows all the different patterns of "wave nonresponse" present among these 5,570 children throughout the four cycles. Wave nonresponse for a given cycle (or wave) means that for that particular cycle the child did not respond. The line corresponding to respondent in cycles 1 and 4 and nonrespondent in 2 and 3 is not included because there are not any cases like that in the dataset. If a kid is nonrespondent for two consecutive cycles, that kid is not sought for interview any longer.

Since there is an interest in examining the factors that contribute to change in aggressive behaviours as children grow, it is important to find what predictors (age, gender, family socio-economic conditions, etc.), and to what degree, affect the development of aggressive behaviours over time. Thomas (2004) found that punitive parenting techniques, age of the kid, the interaction between age and punitive parenting techniques, household income, the interaction between income and age, family structure, region, the interaction between region and age, were all significant in explaining aggressive behaviour and changes in it over time. She

Table 1.1: Patterns of nonresponse and frequencies among the 5,570 children. x means nonrespondent.

Respondent in Cycle 1	Respondent in Cycle 2	Respondent in Cycle 3	Respondent in Cycle 4	Frequency	
Yes	x	Yes	x	71	
Yes	Yes	x	Yes	108	INTERMITTENT
Yes	x	Yes	Yes	150	
-	-	-	-		
Yes	x	x	x	263	
Yes	Yes	x	x	256	MONOTONE
Yes	Yes	Yes	x	557	
Yes	Yes	Yes	Yes	4,165	
				5,570	

also found some borderline evidence that gender may also influence the aggressive behaviour of children. She did not find any evidence for an effect of language of interview. Statistics Canada (2005) also found that income, depression of the person most knowledgeable about the kid (PMK) and the interaction between age and depression of the PMK, are all significantly related to the aggression of the kid. Carrillo et al. (2005) also found that age has a significant effect in explaining children's aggressive behaviours. But they additionally found that the square of age has a significant effect, too; that is, the effect of AGE on PAS is not linear, and a quadratic relation seems to catch that effect better. Whereas the number of hours spent in daycare was borderline.

On these grounds Carrillo-Garcia (2006) examined the following 12 covariates as potentially explanatory for children's aggressive behaviours. A detailed description of these variables can be found in Appendix B. Age, Age^2 (the square of Age), Depression of the PMK, Punitive Parenting Status, Region, GENDER, Family Status, Household Income Status, Hours in Daycare, the Age by Punitive Parenting Interaction, the Age by Household Income Status interaction, and the Age by Region interaction. We will abbreviate these variables as Age, Age^2, DeprePMK, Punitive, Region, GENDER, FamStat, Income, Hours, Age*Puni, Age*Inco, and Age*Regi, respectively. Table 1.2 shows how the 5,570 children break down with respect to the missingness of the outcome variable (PAS) and the 12 covariates across the four cycles.

The naïve approach or "complete case analysis" makes strong assumptions about the response mechanism; it basically assumes that the missing data are missing completely at random (MCAR; see section 2.2.1 for further detail). But this naïve approach is easy to apply and is widely used by practitioners. In complete case analysis, only those subjects who complete all the items of interest in all the cycles are taken into the analysis. So, in principle, for a complete case analysis of the NLSCY we would use the information for all those 4,165 kids in the last row of Table 1.1. However, due to missing items among these kids, only 3,049 can be considered as completers. Among the 4,165 kids there are 81 for which the PAS was not observed in at least one cycle although all their covariates of interest were observed in all four cycles, there are 559 for which at least one covariate was not observed in at least one cycle

Table 1.2: Frequencies of missing PAS and/or covariates among the 5,570 children across the four cycles.

		12 Covariates	
Observed PAS?	Cycle	All observed	One or more missing
Yes	1	5,263	124
	2	4,736	278
	3	4,182	416
	4	3,905	233
No	1	68	115
	2	32	40
	3	14	331
	4	31	254

although their PAS was observed in all four cycles, there are 476 for which the PAS was not observed in at least one cycle and at least one covariate was not observed in at least one cycle, and there are only 3,049 for which the PAS and all the covariates were observed in all cycles. These 3,049 kids are the "completers."

A Note on Weights

For cross-sectional surveys, the survey (or sampling) weight is a number for each element in the sample, which "can be thought of as the number of units in the population represented by the sample member" (Lohr, 1999). These survey weights are usually constructed in several stages. The basic survey weights are the inverse of the probability of inclusion of each unit. They are then adjusted for under-coverage, unit nonresponse, and "calibrated" to known population quantities. For longitudinal surveys, however, the definition is not so straightforward since, in many cases, "the population" changes over time. Therefore, careful specification of the population to which the weights refer is necessary.

Table 1.3 shows five longitudinal weights available in the NLSCY up to cycle 4. All children in the longitudinal sample have positive longitudinal weight for cycle 1. In other words, all longitudinal kids were respondents at the first cycle. This is so, not because there was 100% response rate, but because "the longitudinal sample will be comprised of all children sampled for Cycle 1 of the survey in responding households" (Statistics Canada, 1995). So, in theory, all the kids in our sample should have at least one cycle of data (cycle 1); however, this is not so because of item missing values, as we will see later.

At any given cycle, the only children who have positive values for that cycle's longitudinal weight are those who were respondent at that cycle, irrespective of their response history. The longitudinal weight at any cycle is the longitudinal weight at cycle 1 but adjusted for attrition (and intermittent patterns). All the weights are representing the same population, those kids who were 2-5 years old at 1996. However, if we want to make use of all the

longitudinal children for the analysis, the only available weight which allows this is the first cycle's longitudinal weight.

A different weight which is also available for the NLSCY is the so-called "funnel weight" for cycle 4. This weight is positive only for those children who have responded in all four cycles. Thus using this weight for any given analysis ignores all those kids who have failed to respond in at least one cycle.

Table 1.3: Patterns of the different longitudinal weights available.

Unit	Longitudinal Weight cycle 1	Longitudinal Weight cycle 2	Longitudinal Weight cycle 3	Longitudinal Weight cycle 4	Funnel Weight cycle 4
1	29.7	.	.	.	.
2	30.5	.	.	.	.
3	26.3	23.7	.	.	.
4	1273.0	1862.7	.	.	.
5	16.1	20.4	23.0	.	.
6	46.0	49.1	45.0	.	.
7	58.8	.	50.7	.	.
8	37.8	.	51.7	.	.
9	51.8	54.3	51.7	.	.
10	2069.0	3063.6	3129.2	.	.
11	54.7	63.0	.	49.5	0
12	2002.6	2990.1	.	2509.1	0
13	1237.6	.	1746.5	2392.6	0
14	1688.8	.	1683.0	2925.7	0
15	19.7	21.7	20.3	25.1	23.4
16	3265.9	3924.4	4046.3	4111.5	8910.0

In this thesis we apply modified GEE methods to the NLSCY dataset. We will compare different approaches for handling some of the types of missing patterns found in this dataset (in particular missing responses). One important decision we should make is which weight to use for our analyses. In table 1.3 we can see that if we use any of the longitudinal weights for cycles 2 through 4, we would be ignoring those children who do not respond at that particular cycle. For example, if we use the longitudinal weight for cycle 4, we would not take into account the kids who did not respond at cycle 4. Since this is not an appealing characteristic, because we would like to use as much information as possible, we decide that the most appropriate weight to use is the longitudinal weight for cycle 1. For example, if a kid responded only at cycle 1, we would like to use that cycle's information for that kid; or if a kid responded at cycle 1 and 3, we would like to include those two cycle's information in our analyses. And none of the longitudinal weights from cycle 2 through 4 allow for this.

Similarly, any analysis using the funnel weight for cycle 4 would ignore any cycle's information for any child who failed to respond in at least one cycle, and therefore is not appropriate for our objectives. However, this weight is the most appropriate one to use if one were to use

the naïve "complete case analysis". For complete case analysis, only those subjects who responded in all four cycles (completers), and did not have any item missing, are used. In this case the funnel weight is the most appropriate one to use because Statistics Canada already adjusted these weights in the best possible manner so that the completer kids account for themselves *and* also for the non-completers.

1.3 Brief Summary of Major Results

This thesis develops methods for the analysis of longitudinal surveys when the response variable contains missing values. Our methods are built within the framework of the popular GEE method, which has been extensively studied under the non-survey context.

In chapter 2, we first provide a review of the GEE method and missing data problems discussed in the current literature, and set the stage for our development for longitudinal surveys.

In chapter 3, we consider the use of GEE method for complex survey data with complete responses. We argue that under such scenarios, a joint randomization framework is appropriate for the proposed pseudo-GEE approach. We establish the consistency of the pseudo-GEE estimators under the proposed framework.

Chapter 4 presents methods for the analysis of longitudinal surveys with missing responses. We first extend the weighted GEE method of Robins et al. (1995), described in section 2.2.3, which was proposed under the non-survey context, to handle complex longitudinal surveys. We show how, in addition to the weights used by Robins et al. (1995) (to adjust for non-responses), the survey weights can be incorporated into the analysis.

Our major focus in this chapter, however, is to show that the pseudo-GEE method, discussed in chapter 3, provides valid inferences under hot-deck imputation for missing responses. Both, weighted and unweighted hot-deck imputation methods are considered. Consistency of the resulting pseudo-GEE estimators under imputation is established. Linearization variance estimators are also developed under hot-deck imputation procedures.

In chapter 5, we examine the finite sample performance of the pseudo-GEE estimators as well as the related variance estimators through a simulation study. The finite population structure and variables used in the simulation are tailored from the National Longitudinal Survey of Children and Youth.

We end up this thesis with some conclusions about the methods developed and results obtained, as well as some topics for potential future research in chapter 6.

Chapter 2

The GEE Approach to the Analysis of Non-Survey Data

2.1 The GEE Method for Complete Data

2.1.1 Generalized Linear Models

Generalized Linear Models (GLM) is a method, originally proposed by Nelder and Wedderburn (1972), for estimating the regression between a *univariate* response variable and a set of covariates. This method is of broad application since it generalizes traditional regression models in two ways. Firstly, GLM is suited for situations in which the response variable follows any distribution belonging to the exponential family; and then continuous (linear regression), binary (logistic regression), or count (Poisson regression) outcome variables can be analyzed by GLM. Regression analysis, on the other hand, is applicable only when the (continuous) response variable follows a normal distribution. Additionally, whereas regression assumes the expected value of the response is a linear combination of the covariates (linear predictor), in GLM this expected value is allowed to be *some suitable function* of the linear predictor. This is much less restrictive.

Hardin and Hilbe (2001) characterize the GLM as being composed of the following items: a random component for the response Y, following an exponential family distribution; a systematic component, specifying that the effect of the covariates X on the mean of Y can be expressed by way of the "linear predictor" $\eta = X'\boldsymbol{\beta}$; a known monotonic, one-to-one, differentiable "link function" $g(\cdot)$ relating the mean of the response Y to the linear predictor; and that the variance of the response may change with the covariates only as a function of the mean.

We now give a more thorough explanation of GLM (following the lines of Fitzmaurice et al., 2004), which will later serve as the basis for the presentation of the main method used in this thesis, namely GEE. We suppose that we have a random sample of n independent observations of a variable Y, with distribution usually assumed to belong to the exponential family, denoted by $Y_1, Y_2, \cdots, Y_n$, and with expected value $E[Y_i] = \mu_i$. It is possible to consider other types of distributions for the GLM method; for instance, dispersion models, as described

in Song (2007). Additionally, associated with each individual i we also observe a set of p fixed explanatory variables $X_{i1}, X_{i2}, \cdots, X_{ip}$. We set the, n, Y_i values in a column vector Y of dimension $n \times 1$; all the X values for individual i in a $p \times 1$ vector X_i; and all the X_i vectors in a $p \times n$ matrix X, as

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{pmatrix}, \quad X = (X_1, X_2, \ldots, X_n) = \begin{pmatrix} X_{11} & X_{21} & \ldots & X_{n1} \\ X_{12} & X_{22} & \ldots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \ldots & X_{np} \end{pmatrix}.$$

The random component of the model specifies that the variable Y_i follows an exponential family distribution, whose probability density (or mass) function can be written as

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}, \tag{2.1}$$

where $E[Y_i] = \mu_i = db(\theta_i)/d\theta$, θ_i is called the "canonical location parameter," and $\phi > 0$ is called the "dispersion parameter." We also obtain that $\text{Var}[Y_i] = \phi\, d^2 b(\theta_i)/d\theta^2 = \phi d\mu/d\theta = \phi \upsilon(\mu_i)$, where $\upsilon(\mu_i) = d^2 b(\theta_i)/d\theta^2 = d\mu_i/d\theta$ is called the "variance function," a known function of the mean μ_i. Hardin and Hilbe (2001) also point out that in GLM one assumes that the variance function $\upsilon(\mu_i)$ and the dispersion parameter ϕ are correctly specified.

The systematic component of the model determines that the effect of the covariates X_i on μ_i can be expressed in terms of the linear predictor $\eta_i = X_i' \boldsymbol{\beta} = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$. Where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)'$ is a $p \times 1$ vector of unknown regression coefficients. It is assumed that this linear predictor is correctly specified (Hardin and Hilbe, 2001).

The link function is a known one-to-one transformation $g(\cdot)$ which relates μ_i to the linear predictor η_i as $g(\mu_i) = \eta_i = X_i' \boldsymbol{\beta} = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$; or equivalently, $\mu_i = g^{-1}(\eta_i)$. The function $g(\cdot)$ could be any, but the most common choice is the "canonical link function" which transforms μ_i to the canonical location parameter. The link function has the purpose of transforming the range of possible values of the linear predictor to the range of possible values of the response variable. Again, it is assumed that this link function is correctly specified.

As examples of GLMs we can mention the ones obtained for the following random components: the Normal distribution, for which the variance function is $\upsilon(\mu_i) = 1$, in this case the canonical link is the identity ($\mu_i = \eta_i$), and $\phi = \sigma^2$, the variance of the distribution; the Bernoulli distribution, for which the variance function is $\upsilon(\mu_i) = \mu_i(1 - \mu_i)$, the canonical link is the "logit" ($\log(\mu_i/(1-\mu_i)) = \eta_i$), and $\phi = 1$; and the Poisson distribution, for which $\upsilon(\mu_i) = \mu_i$, the canonical link is the natural logarithm ($\log(\mu_i) = \eta_i$), and $\phi = 1$. The first case leads to common linear regression, the second is known as logistic regression, and the last one as Poisson regression.

Once an exponential family distribution has been appropriately selected for the response, one chooses explanatory variables and a suitable link function to match the mean response to the linear predictor, and then the estimation of the regression coefficients $\boldsymbol{\beta}$ follows. The most common method of estimation is the method of maximum likelihood (ML). It has the intuitive interpretation that the estimator obtained (MLE) is the value of the parameter which is most

likely (probable) to be the true one with the data at hand. Other important properties of the MLE are that they are consistent, asymptotically efficient and normally distributed. A proof of the consistency can be found in pp. 444-445, and the proof of asymptotic normality and efficiency in pp. 449-450, in Lehmann and Casella (1998).

To get the MLE of $\boldsymbol{\beta}$ we need to determine the likelihood function of $Y = (Y_1, Y_2, \cdots, Y_n)'$. Since GLM assumes that the n observations are independent, the likelihood function is the product of the probability density (or mass) functions of the n single observations, 2.1. The likelihood function is then

$$L = L(\theta, \phi; y_i, y_2, \cdots, y_n) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}.$$

To find the MLE of $\boldsymbol{\beta}$ we need to maximize L with respect to $\boldsymbol{\beta}$, which is equivalent to maximizing the log-likelihood

$$l = \log(L) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}$$

with respect to $\boldsymbol{\beta}$. We take the derivative of l with respect to $\boldsymbol{\beta}$ (using the chain rule), set it equal to zero and solve for $\boldsymbol{\beta}$:

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\} \\ &= \sum_{i=1}^{n} \left[\frac{\partial}{\partial \theta_i} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\} \left(\frac{\partial \theta_i}{\partial \mu_i} \right) \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right] = \mathbf{0}, \end{aligned}$$

which, since $db(\theta_i)/d\theta = \mu_i$ and $d\mu_i/d\theta = \upsilon(\mu_i)$, becomes

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{y_i - db(\theta_i)/d\theta}{\phi} \left(\frac{\partial \theta_i}{\partial \mu_i} \right) \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right] = \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{\phi} \frac{1}{\upsilon(\mu_i)} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right] = \mathbf{0}.$$

For a motive that will become clear in the next section, we write this last "estimating equation" as

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \left[\phi\, \upsilon(\mu_i) \right]^{-1} (y_i - \mu_i) = \mathbf{0}; \tag{2.2}$$

where $\partial \mu_i / \partial \boldsymbol{\beta}$ is the $p \times 1$ vector of derivatives of μ_i with respect to each of the β_j's ($\partial \mu_i / \partial \boldsymbol{\beta} = (\partial \mu_i / \partial \beta_1, \partial \mu_i / \partial \beta_2, \cdots, \partial \mu_i / \partial \beta_p)'$) and the right-hand side is a $p \times 1$ vector of zeroes. Equation 2.2 is then solved for $\boldsymbol{\beta}$ to find the MLE. In some cases this may require an iterative procedure like Newton-Raphson; for example in the cases of logistic and Poisson regression. The solution $\hat{\boldsymbol{\beta}}$ obtained by this procedure is consistent for $\boldsymbol{\beta}$ with the only requirement that the linear predictor and link function be correctly specified; but one should use robust estimators of the variance of $\hat{\boldsymbol{\beta}}$ whenever the variance of Y_i may be misspecified (Fitzmaurice et al., 2004).

As we discussed earlier, one of the assumptions of GLM is that the n observations form an *independent* sample from the population. However, in longitudinal studies the observations are not independent altogether. Observations from the same individual (over time) generally embody some correlation even if the individuals are independent of one another. In the next section we present a method for estimating regression coefficients from different models which allows for correlation among observations, and is an extension of GLM.

2.1.2 Generalized Estimating Equations

The method of Generalized Estimating Equations (GEE) was proposed in a seminal paper by Liang and Zeger (1986). This method is applicable to clustered data in general, and longitudinal studies in particular; it permits estimation of regression coefficients in the presence of within subject correlation arising in this kind of studies.

In the previous section we found the estimating equation for obtaining the MLE of the regression coefficients $\boldsymbol{\beta}$. For this task it is necessary to use the likelihood function for the data. In that case it was simple to get given that all the observations were independent and it was simply the product of the n probability density (mass) functions for the single observations. In longitudinal studies, however, observations coming from the same individual are not independent and therefore the likelihood function of the data is not simply the product of the single probability density (mass) functions. So, if we wanted to find the MLE of some regression coefficients in a longitudinal study, we would need to posit a *joint* probability density (mass) function for the responses coming from a single individual. This is not a simple task in general. For example, for continuous, normally distributed, responses we can, relatively easily, posit such a joint distribution for the responses from the same individual; but for discrete responses, such as binary or count outcomes, it is not simple at all. The method of GEE is an attempt to get estimators without the requirement of assuming a fully parametric distribution for the response, but only a regression model for its mean. Thus this method does not produce MLEs.

We now present a detailed explanation of GEE (following the lines of Fitzmaurice et al., 2004). We assume that we have an independent random sample of n subjects. For each subject i, we take a set of T_i repeated measurements (over time) of a random variable Y, our outcome of interest. We denote these T_i measurements for individual i by $Y_{i1}, Y_{i2}, \cdots, Y_{iT_i}$, or Y_{ij}, $j = 1, 2, ..., T_i$; and we set their expected value to $E[Y_{ij}] = \mu_{ij}$. Additionally, associated with each observation Y_{ij} we also observe a set of p explanatory variables $X_{ij1}, X_{ij2}, \cdots, X_{ijp}$, or X_{ijk}, $k = 1, 2, ..., p$. We set all the Y_{ij} values for subject i in a column vector Y_i of dimension $T_i \times 1$, all the X_{ijk} values for individual i at time j in a $p \times 1$ vector X_{ij}, and all the X'_{ij} vectors for subject i in a $T_i \times p$ matrix X_i, as

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iT_i} \end{pmatrix}, \; X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \; X_i' = (X_{i1}, X_{i2}, \ldots, X_{iT_i})' = \begin{pmatrix} X_{i11} & X_{i12} & \ldots & X_{i1p} \\ X_{i21} & X_{i22} & \ldots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{iT_i1} & X_{iT_i2} & \ldots & X_{iT_ip} \end{pmatrix}. \tag{2.3}$$

With this notation we have that $E[Y_i] = \boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{iT_i})'$, and since we assume that the

observations between subjects are independent, we have that the, n, Y_i vectors are independent; although the T_i elements in vector Y_i are not independent from one another. In fact, we assume that there exists some general variance-covariance matrix of Y_i denoted by Σ_i. In GEE we are interested in modeling the mean response $\boldsymbol{\mu}_i$ but not the covariance matrix Σ_i; we regard this matrix as a nuisance parameter but include it in the model just to account for the autocorrelation within subjects.

The GEE method can be characterized as being composed of the following four items. 1. A systematic component, specifying that the effect of the covariates X_{ij} on the mean of Y_{ij}, μ_{ij}, can be expressed by way of the "linear predictor" $\eta_{ij} = X'_{ij}\boldsymbol{\beta} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)'$ is a $p \times 1$ vector of unknown regression coefficients. 2. A known monotonic, one-to-one, differentiable "link function" $g(\cdot)$ relating the mean of the response to the linear predictor, as $g(\mu_{ij}) = \eta_{ij} = X'_{ij}\boldsymbol{\beta}$; or equivalently, $\mu_{ij} = g^{-1}(\eta_{ij})$. The link function has the purpose of transforming the range of possible values of the linear predictor to the range of possible values of the response variable. 3. The variance of the response Y_{ij}, given X_{ij}, may change with X_{ij} only as a function of the mean, μ_{ij}, as $\text{Var}[Y_{ij}] = \phi\,\upsilon(\mu_{ij})$; where $\phi > 0$ is called the "dispersion parameter," and $\upsilon(\mu_{ij})$ is called the "variance function," a known function of the mean μ_{ij}. 4. The within-subject association for the repeated measurements is assumed to be a function of the means μ_{ij} and some additional parameters α, which we do not model. We proceed as follows. We pose a "working" correlation matrix, $\mathbf{R}_i(\alpha)$, for Y_i, depending on some parameters α which we estimate from the data. Therefore, the "working" variance-covariance matrix of Y_i is composed as $V_i = A_i^{1/2}\mathbf{R}_i(\alpha)A_i^{1/2}$; where A_i is a $T_i \times T_i$ diagonal matrix with $\text{Var}[Y_{ij}] = \phi\,\upsilon(\mu_{ij})$ as the jth diagonal element. If we happen to specify $\mathbf{R}_i(\alpha)$ to be the true correlation matrix of Y_i, then $V_i = \Sigma_i = \text{Cov}[Y_i]$.

Items 1 through 3 in the previous characterization of GEE are equivalent to those assumed in GLM. However, whereas in GLM we posit a fully parametric model for the distribution of the observed responses Y_i, in GEE we replace that requirement by the much weaker specification of only the first (and second) moments of the vector Y_i. "It is the fourth component, the incorporation of the within-subject association among the repeated responses from the same individual, that represents the main extension of GLM to longitudinal data" (Fitzmaurice et al., 2004).

As examples of GEEs we can mention the following: a continuous response with $\upsilon(\mu_{ij}) = 1$, the identity link ($\mu_{ij} = \eta_{ij}$), ϕ is a variance term, and $\mathbf{R}_i(\alpha)_{[j,k]} = \alpha^{|k-j|}$ or AR(1) structure, this is a linear regression with AR(1) correlation structure, though any other correlation structure could be used if appropriate; a binary response with $\upsilon(\mu_{ij}) = \mu_{ij}(1-\mu_{ij})$, with the logit link ($\log(\mu_{ij}/(1-\mu_{ij})) = \eta_{ij}$), $\phi = 1$, and unstructured association, this is a logistic regression model with Bernoulli variance assumption and unstructured association; and if the response outcome is a count, with $\upsilon(\mu_{ij}) = \mu_{ij}$, and the natural logarithm as link function ($\log(\mu_{ij}) = \eta_{ij}$), ϕ is an added variance term (overdispersion beyond the usual Poisson's variance), and unstructured association, this is a log-linear regression model with overdispersion and unstructured association.

Once explanatory variables and a suitable link function to match the mean response to the linear predictor are selected, one should also choose a variance function and a dispersion parameter. This should be based on the nature of the response; for example if the response is

binary, a suitable variance function is $v(\mu_{ij}) = \mu_{ij}(1-\mu_{ij})$ and a suitable dispersion parameter is $\phi = 1$. Then, also, a within-subject association structure should be chosen. This can be based on subject matter knowledge, sometimes it can be a by-product of the nature of the response, or could be unspecified if we lack knowledge about it and/or there is a large sample size. After the model is set up then the estimation of the regression coefficients $\boldsymbol{\beta}$ can take place. The GEE estimator $\hat{\boldsymbol{\beta}}$ is obtained as a solution to the following set of equations:

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{2.4}$$

where $\partial \boldsymbol{\mu}_i'/\partial \boldsymbol{\beta}$ is the $p \times T_i$ matrix of partial derivatives of (the vector) $\boldsymbol{\mu}_i'$ with respect to each of the β_j's:

$$\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial \mu_{i1}}{\partial \boldsymbol{\beta}} & \frac{\partial \mu_{i2}}{\partial \boldsymbol{\beta}} & \cdots & \frac{\partial \mu_{iT_i}}{\partial \boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \partial\mu_{i1}/\partial\beta_1 & \partial\mu_{i2}/\partial\beta_1 & \cdots & \partial\mu_{iT_i}/\partial\beta_1 \\ \partial\mu_{i1}/\partial\beta_2 & \partial\mu_{i2}/\partial\beta_2 & \cdots & \partial\mu_{iT_i}/\partial\beta_2 \\ \vdots & \vdots & \ddots & \vdots \\ \partial\mu_{i1}/\partial\beta_p & \partial\mu_{i2}/\partial\beta_p & \cdots & \partial\mu_{iT_i}/\partial\beta_p \end{pmatrix},$$

and the right-hand side is a $p \times 1$ vector of zeroes. Equation 2.4 for GEE has a similar form to equation 2.2 for GLM.

The left-hand side of equation 2.4 is a function of $\boldsymbol{\beta}$ and α, and in general the equation does not have a closed form solution, but instead has to be solved iteratively. This iterative procedure can be summarized in the following four steps. 0. Obtain an initial estimate of $\boldsymbol{\beta}$. 1. Using the current estimate $\boldsymbol{\beta}_{(l)}$ calculate the standardized residuals, and get estimates of α, (ϕ if required,) $\mathbf{R}_i(\alpha)$, and V_i. 2. Using the current values of $\boldsymbol{\beta}$, V_i, and $\boldsymbol{\mu}_i$, update the estimate of $\boldsymbol{\beta}$, using Fisher scoring algorithm, by

$$\boldsymbol{\beta}_{(l+1)} = \boldsymbol{\beta}_{(l)} + \left[\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_{(l)}} V_{i(l)}^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}_{(l)}}\right]^{-1} \left[\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_{(l)}} V_{i(l)}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i(l)})\right].$$

3. Iterate steps 1 and 2 until convergence. The standardized residuals are given by

$$e_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{v(\hat{\mu}_{ij})}},$$

where $\hat{\mu}_{ij} = g^{-1}\left(X_{ij}'\hat{\boldsymbol{\beta}}\right)$; the dispersion parameter ϕ is estimated by

$$\hat{\phi} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{T_i} e_{ij}^2}{(\sum_{i=1}^{n} T_i) - p};$$

and α is estimated according to the assumed autocorrelation structure. We will assume that all the subjects share a common within-subject association[1] and we will not specify any structure

[1] Although for the binary response case to be studied in chapter 5, the within-subject association depends on $\boldsymbol{\mu}_i$ for each subject i.

for it. In other words, we assume that the working within-subject correlation matrix is the same for all individuals and we do not constraint this single matrix in any way. We can write these assumptions, mathematically, as:

$$\mathbf{R}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha_{jk} & j \neq k; \end{cases}$$

and we estimate it with:

$$\hat{\alpha}_{jk} = \frac{\sum_{i=1}^{n} e_{ij} e_{ik}}{(n-p)\hat{\phi}}.$$

This procedure is implemented in several statistical packages for different outcome variables and within-subject correlation structures, for example R (function `gee`) and SAS (procedure `genmod`).

The GEE estimator $\hat{\boldsymbol{\beta}}$, obtained by the procedure outlined, is consistent for $\boldsymbol{\beta}$, with the only requirement that the linear predictor and link function be correctly specified. The consistency of $\hat{\boldsymbol{\beta}}$ does not depend on the validity of the assumed correlation matrix $\mathbf{R}_i$; however, $\hat{\boldsymbol{\beta}}$ will be more efficient if $\mathbf{R}_i$ resembles Σ_i more closely. One should use robust estimators of the variance of $\hat{\boldsymbol{\beta}}$ whenever $\mathbf{R}_i$ may be misspecified. We also have, asymptotically, $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathbf{MVN}(\mathbf{0}, \mathrm{Cov}(\hat{\boldsymbol{\beta}}))$, where

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = n\Big[\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\Big]^{-1} \Big[\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \Sigma_i V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\Big] \Big[\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\Big]^{-1}; \qquad (2.5)$$

which reduces to $n[\sum_{i=1}^{n} {}^{\partial \boldsymbol{\mu}_i'}/_{\partial \boldsymbol{\beta}} V_i^{-1} {}^{\partial \boldsymbol{\mu}_i}/_{\partial \boldsymbol{\beta}}]^{-1}$ if $V_i = \Sigma_i$. $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ can be estimated by replacing the values of α, ϕ, and $\boldsymbol{\beta}$ in 2.5 by their estimated values $\hat{\alpha}$, $\hat{\phi}$, and $\hat{\boldsymbol{\beta}}$, and Σ_i by $(y_i - \hat{\boldsymbol{\mu}}_i)(y_i - \hat{\boldsymbol{\mu}}_i)'$. This variance is also obtained in the software packages R (`gee`) and SAS (`genmod`).

We end up this section giving a short motivation for the estimating equations 2.4 in GEE. In generalized least squares estimation for linear regression, the vector $\boldsymbol{\beta}$ which minimizes

$$\sum [y_i - X_i\boldsymbol{\beta}]' \Sigma_i^{-1} [y_i - X_i\boldsymbol{\beta}]$$

solves the equation $\sum X_i' \Sigma_i^{-1} (y_i - X_i\boldsymbol{\beta}) = \mathbf{0}$. In GEE, a vector $\boldsymbol{\beta}$ which minimizes

$$\sum [\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})]' V_i^{-1} [\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})]$$

also solves equation 2.4. So that GEE can also be thought of as a generalization to longitudinal responses of the generalized least squares technique.

2.2 The GEE Method for Incomplete Data

2.2.1 General Nonresponse Mechanisms

Here we briefly review the three general mechanisms of missing values and in the next section we will elaborate on this discussion for longitudinal studies. The theoretical layout for

missing mechanisms was set up by Rubin (1976). In this section we assume that we have a matrix $Y_{n\times p}$ of p intended responses for n subjects. However, due to nonresponse, some of the components of this matrix are missing. We denote by Y_{obs} the components of matrix Y which are actually observed, and by Y_{mis} the missing components of Y. So that $Y = (Y_{\text{obs}}, Y_{\text{mis}})$. Additionally, we use a matrix $R_{n\times p}$, called the response indicator matrix, to indicate whether the corresponding element in Y is observed or not; we construct this matrix as follows. Element ij in matrix R is set equal to one if element ij is observed and equal to zero if not. We characterize the nonresponse mechanisms, as in Little and Rubin (2002), by the conditional distribution of R given Y (the variable of interest) and X (the set of covariates); we denote this distribution by $f(R|Y,X,\xi)$, where ξ denotes some unknown parameters which define the response/nonresponse mechanism.

In lay terms, the response mechanism is said to be missing completely at random (MCAR) if missingness does not depend on the actual values of the intended response. In other words, MCAR means that the probability that an element of R is equal to one is independent of Y. This can be written more formally as

$$f(R|Y,X,\xi) = f(R|\xi).$$

The response mechanism is said to be missing at random (MAR) if the missingness, conditionally, does not depend on any *unobserved* values, though it may depend on observed values. In other words, the response mechanism is MAR if the probability that an element of R is equal to one is independent of Y_{mis}. We write this definition technically as

$$f(R|Y_{\text{obs}}, Y_{\text{mis}}, X, \xi) = f(R|Y_{\text{obs}}, X, \xi).$$

Note that if the response mechanism is MCAR it is also MAR. In other words, the MAR mechanism is more general than MCAR.

The third response mechanism is called not missing at random (NMAR), and it means that the missingness depends on the unobserved values. In this case, the probability that an element of R is equal to one depends on Y_{mis}. Or equivalently, the conditional distribution of R given Y, $f(R|Y_{\text{obs}}, Y_{\text{mis}}, X, \xi)$, cannot be simplified any further since it depends on some components of Y_{mis}. In NMAR situations the response mechanism cannot be ignored for drawing inferences; this is why it is also called "nonignorable". In contrast to MCAR and MAR, which are called "ignorable."

In practice, NMAR would be a safer assumption, but it is also an assumption which is difficult to deal with for data analysis. And indeed, any one of the three missing mechanisms is difficult to verify. A common practice is to assume MAR. If the (main) model is reasonable, and if there exists a rich set of covariates which is observed, then the MAR assumption would be close to an NMAR, if major covariates are used in the model.

2.2.2 Nonresponse in Longitudinal Studies

In longitudinal studies the nonresponse patterns and mechanisms get more complicated. We intend to measure a response variable and a set of covariates on each of n subjects at each of,

say, T times (cycles). However, generally a variety of different missing patterns occur and we cannot observe all the intended measures. Table 2.1 shows some of the patterns that can appear with two covariates and three cycles.

In longitudinal studies, some subjects respond at all cycles and to all variables of interest. These units are called complete cases. Unit 1 in table 2.1 is an example of a complete case. Some units fail to respond altogether, at all cycles and all variables of interest. This situation is referred to as unit (or total) nonresponse (Statistics Canada, 2003). Unit 10 in table 2.1 is an example of unit nonresponse. It is usually dealt with by reweighing the respondent units to account for these unit-nonrespondents. For a comprehensive discussion of weighting for non response see, for example, Kalton and Kasprzyk (1986), Beaumont (2005), and Little and Vartivarian (2005).

Another situation arises when a subject is observed for some cycles but not for others. Those cycles in which a subject is not observed at all are sometimes referred to as wave nonresponse. In table 2.1, cycle 3 for unit 4, cycle 1 for unit 6, cycle 2 for units 8 and 9, and cycles 2 and 3 for unit 7 are examples of wave nonresponse. In the NLSCY this kind of nonresponse is handled cycle by cycle. Each cycle's longitudinal weight is adjusted so that the wave respondents for that cycle account for the wave nonrespondents and the weight for the latter are set to zero.

For some units, it may happen that, at some cycles, some of the covariates are not observed, whereas the response variable is observed. This situation is sometimes referred to as missing covariates, missing X's, or missing regressors (Little, 1992; Parzen et al., 2002; Chen, 2004). Some examples of this case, in table 2.1, are cycle 1 for units 4, 5, and 9. Unit 2 is an extreme example of missing covariates, in which the outcome variable Y is observed at all cycles whereas all the covariates are unobserved at all cycles.

At some cycles, some units may have the response variable Y not observed, whereas all the covariates are observed. This situation is sometimes referred to as missing outcome or missing response (see Rotnitzky et al. 1997, Wang et al. 2004, for analysis in the ignorable case, and Baker and Laird 1988, Fitzmaurice et al. 1996, Rotnitzky et al. 1998, in the nonignorable case). Some examples of this case, in table 2.1, are cycle 2 for unit 4 and cycle 3 for unit 9. Unit 3 is an extreme example of missing outcome, for which the outcome variable Y is never observed at any cycle whereas all the covariates are always observed.

There are also some units which, at some cycles, have missing outcomes and missing covariates at the same time. For example, in table 2.1, units 5 and 6 have missing outcome and missing covariates at cycles 2 and 3.

Additionally, in longitudinal studies it often occurs that when a subject misses one wave, that subject never returns to the study. In other words, once there is a wave nonresponse for some unit, that unit is likely to have wave nonresponse from then on. These units are called dropouts (Fitzmaurice et al., 1995; Preisser et al., 2000; Yi and Thompson, 2005). Unit 7 in table 2.1 is an example of a dropout at cycle 2, and unit 4 is an example of dropout at cycle 3. On the other hand, in some cases, some units who miss a wave may come back to the study at a later wave. These units can be called intermittent observations or units (see Fitzmaurice et al. 2004, sec. 14.4; Robins et al. 1995). Examples of intermittent observations, in table 2.1, are units 6, 8, and 9. Unit 6 missed cycle 1 but came back at cycle 2, units 8 and 9 missed cycle 2

Table 2.1: Some response patterns occurring in longitudinal studies. "Yes" means observed and x means not observed.

unit	cycle	response Y	covariate X_1	covariate X_2
1	1	Yes	Yes	Yes
	2	Yes	Yes	Yes
	3	Yes	Yes	Yes
2	1	Yes	x	x
	2	Yes	x	x
	3	Yes	x	x
3	1	x	Yes	Yes
	2	x	Yes	Yes
	3	x	Yes	Yes
4	1	Yes	x	x
	2	x	Yes	Yes
	3	x	x	x
5	1	Yes	x	x
	2	x	Yes	x
	3	x	x	Yes
6	1	x	x	x
	2	x	Yes	x
	3	x	x	Yes
7	1	Yes	Yes	Yes
	2	x	x	x
	3	x	x	x
8	1	Yes	Yes	Yes
	2	x	x	x
	3	Yes	Yes	Yes
9	1	Yes	x	Yes
	2	x	x	x
	3	x	Yes	Yes
10	1	x	x	x
	2	x	x	x
	3	x	x	x

but came back at cycle 3. The set consisting of all complete observations and dropouts is called monotone dataset. We call the monotone dataset together with the intermittent observations the intermittent dataset.

We now discuss the three general missing data mechanisms introduced in section 2.2.1 for the case of wave nonresponse. A dropout mechanism is missing completely at random (MCAR) if it is independent of the measurement process. In other words, dropout is MCAR if the probability of dropout at any given wave is independent of all observed (past) and unob-

served (present and future) outcomes. The dropout mechanism is missing at random (MAR) if it depends on the past (observed) outcomes but is independent of the current (missing) and future (missing) outcomes. And the dropout mechanism is not missing at random (NMAR) if it depends on the actual values of those unobserved outcomes from dropout onward. Fitzmaurice et al. (2004) explain these mechanisms along these lines; with MAR, the people who dropout at time t and those who remain in the study have the same distribution of present and future observations conditional on them having the same past; with NMAR those two distributions are different, even after conditioning on the past history. MCAR means that at any time, those who drop out have the same distribution of present and future observations as those who remain in the study.

Intermittent patterns are a bit more complicated. We say the intermittent missing mechanism is missing completely at random (MCAR) if it is independent of the measurement process. This is, if the probability of missingness at a given wave is independent of all observed and unobserved outcomes either in the past, present, or future waves. The intermittent missing mechanism is missing at random (MAR) if it is independent of the unobserved portion of the measurement process. This is, if at a given wave, the probability of nonresponse depends on the observed outcomes (either past or future) but is independent of the unobserved ones. And the intermittent missing mechanism is not missing at random (NMAR) if it depends on the unobserved portion of the measurement process. In other words, if at a given wave, the probability of nonresponse depends on any unobserved outcomes (past, present, or future).

The GEE analysis using only the complete cases, i.e. those subjects with all waves completely observed, is valid only under the strongest assumption of MCAR. Here we use the word "valid" to indicate that the regression coefficients obtained with this approach are consistent. This approach is widely used because it is readily applicable using standard complete case GEE software. However, even if the nonresponse mechanism *is* MCAR, complete case analyses are usually inefficient (high variances) because of the reduction in number of observations compared to an analysis which uses all the observed waves for each subject. We call this approach "available case" (AC) analysis. This method is more efficient than a complete case analysis because it uses more data; but nonetheless, it yields consistent estimators of the regression coefficients only under a MCAR mechanism (Albert, 1999), just like complete case analysis. In the next subsection we describe a modified GEE method valid under the less stringent assumption of waves MAR.

2.2.3 The Weighted GEE Using Response Probabilities

Monotone Missingness

Either complete case or available case GEE analysis with either monotone or intermittent missing data produces consistent estimates of the regression coefficients only in the case of MCAR data. Robins et al. (1995) propose an extension of the GEE method, applicable to longitudinal studies with missing observations, when the missing mechanism is MAR.

Diggle et al. (2002) summarize the idea of this method, for monotone datasets, along the following lines. If p_{ij} is the probability that subject i has not dropped out by time j, given

his/her observed history, then (under MAR) the observation y_{ij} is representative of all subjects who do drop out and have the same history; therefore, in an available case GEE analysis, the contribution of y_{ij} needs to be weighted by the inverse of p_{ij}, to account for those who dropped out and have the same history. This methodology is sometimes called weighted generalized estimating equations (WGEE).

The WGEE is well suited for our NLSCY dataset because it "requires, inevitably, that we can consistently estimate the dropout probabilities for each subject given their observed measurement history and any relevant covariates. This makes the method best suited to large-scale studies" (Diggle et al., 2002). We next give a detailed explanation of the WGEE methodology for monotone longitudinal datasets. We do not follow the exact same notation used in that paper; we rather adopt the notation proposed by Hardin and Hilbe (2003), which is more consistent with the notation in this thesis.

For this method we assume that the population satisfies the model

$$\xi : \begin{cases} E[Y_{ij}|X_{ij}] = \mu_{ij} = g^{-1}(\eta_{ij}) = g^{-1}(X'_{ij}\boldsymbol{\beta}); \quad i = 1,2,\ldots;\ j = 1,2,\ldots \\ \text{Var}[Y_{ij}|X_{ij}] = \phi\, \upsilon(\mu_{ij}); \quad i = 1,2,\ldots;\ j = 1,2,\ldots \\ \text{Cov}[Y_i|X_i] = A_i^{1/2}\mathbf{R}_i(\alpha)A_i^{1/2}; \quad A_i = \text{diag}[\phi\, \upsilon(\mu_{ij})];\ \mathbf{R}(\alpha) = \text{working correlation matrix} \\ \forall k \neq l,\ Y_k = (Y_{k1}, Y_{k2}, \cdots) \text{ and } Y_l = (Y_{l1}, Y_{l2}, \cdots) \text{ are independent vectors given } X_k, X_l. \end{cases}$$

We select n independent subjects from the population. We intend to measure each subject, i, T times, but some subjects fail to respond after a given cycle and so, drop out of the study. We intend to measure Y_i and X'_i, as in equations 2.3, for all subjects, but due to dropouts the full vector Y_i and matrix X'_i are not always observed. We assume that in addition to Y_{it} and X_{it}, at time t we also measure a vector of covariates V_{it}, $t = 1,2,...,T$. These extra variables V are used to help us understand better the nonresponse mechanism, but are not included in the main model ξ.

We define the response indicator variable R_{it} in the following way,

$$R_{it} = \begin{cases} 1 & \text{if} \quad \text{subject } i \text{ is observed at time } t \\ 0 & \text{otherwise.} \end{cases}$$

We assume that at any give time t, Y_{it}, X_{it}, and V_{it} are either all observed or all missing. In this subsection we only deal with dropouts; i.e. if $R_{it} = 0$ then $R_{i(t+1)} = 0$. Additionally, this method assumes that all subjects are observed at wave 1; i.e. $R_{i1} = 1$ for all subjects.

The MAR assumption, under these settings, can be written as

$$\begin{aligned} P(R_{it} = 1|R_{i(t-1)} = 1,\, &X_{i1},...,X_{iT},\, V_{i1},...,V_{iT},\, Y_{i1},...,Y_{iT}) \\ =&P(R_{it} = 1|R_{i(t-1)} = 1,\, X_{i1},...,X_{i(t-1)},\, V_{i1},...,V_{i(t-1)},\, Y_{i1},...,Y_{i(t-1)}). \end{aligned}$$

This means that, among those subjects observed at time $t-1$, the probability of being observed at time t depends on the past (observed) measurements, but is independent of the current and future measurements. The MAR mechanism implies the following, weaker, assumption; which is sufficient for the WGEE method to be valid;

$$P(R_{it} = 1|R_{i(t-1)} = 1,\, X_{i1},...,X_{i(t-1)},\, V_{i1},...,V_{i(t-1)},\, Y_{i1},...,Y_{iT})$$

$$=P(R_{it}=1|R_{i(t-1)}=1,\, X_{i1},...,X_{i(t-1)},\, V_{i1},...,V_{i(t-1)},\, Y_{i1},...,Y_{i(t-1)}) = p_{it} \qquad (2.6)$$

In words of Robins et al. (1995), this last assumption means that, among subjects observed at time $t-1$, nonresponse at time t is unrelated to the current and future outcomes $Y_{it},...,Y_{iT}$, conditional on the observed past $X_{i1},...,X_{i(t-1)}$, $V_{i1},...,V_{i(t-1)}$, $Y_{i1},...,Y_{i(t-1)}$. We assume that the probability p_{it} of being observed at time t, having been observed at time $t-1$ (conditional on $X_{i1},...,X_{i(t-1)}$, $V_{i1},...,V_{i(t-1)}$, $Y_{i1},...,Y_{i(t-1)}$), is bigger than zero for all times $t=2,...,T$. When the MAR assumption does not hold, the missing mechanism is nonignorable, and the method described here does not work; Rotnitzky et al. (1998) propse a suitable extension in those conditions.

Assumption 2.6 is not testable because it depends on some unobservable quantities ($Y_{it},...,Y_{iT}$). Therefore it is necessary to include as many variables as possible in V_{it}, for all times t in which subject i is observed, in order to ensure that equation 2.6 holds, at least to a good approximation (Robins et al., 1995).

We assume that the response probabilities p_{it} are a known function (taking values on $[0,1]$) of an unknown parameter λ, and the observed past $X_{i1},...,X_{i(t-1)}$, $V_{i1},...,V_{i(t-1)}$, $Y_{i1},...,Y_{i(t-1)}$; i.e. $p_{it}=p_{it}(X_{i1},...,X_{i(t-1)},\, V_{i1},...,V_{i(t-1)},\, Y_{i1},...,Y_{i(t-1)};\, \lambda)$. We will assume that p_{it} follows the logistic regression model

$$\text{logit}(p_{it}) = \log\left(\frac{p_{it}}{1-p_{it}}\right) = \lambda' h(X_{i1},...,X_{i(t-1)},\, V_{i1},...,V_{i(t-1)},\, Y_{i1},...,Y_{i(t-1)}) \qquad (2.7)$$

where $\lambda' = (\lambda_1,...,\lambda_q)$, and $h(\cdot)$ is a $q\times 1$ known vector function of $X_{i1},...,X_{i(t-1)}$, $V_{i1},...,V_{i(t-1)}$, $Y_{i1},...,Y_{i(t-1)}$. Robins et al. (1995) also argue that "standard procedures can be used to investigate the functional form of $p_{it}(\cdot)$, [and...] augmenting the model for p_{it} will usually lead to an improvement of the efficiency with which we estimate $\boldsymbol{\beta}$."

Now we let $\hat{\lambda}$ be the partial pseudo maximum likelihood estimator of λ; i.e. $\hat{\lambda}$ maximizes the partial pseudo log-likelihood

$$l(\lambda) = \sum_{i=1}^{n} l_i(\lambda) = \sum_{i=1}^{n}\sum_{t=1}^{T} \log\left\{p_{it}(\lambda)^{R_{it}}[1-p_{it}(\lambda)]^{1-R_{it}}\right\}^{R_{i(t-1)}}. \qquad (2.8)$$

And we define $\pi_{it}(\lambda) = p_{i1}(\lambda)\times\cdots\times p_{it}(\lambda)$; which, under MAR, is the probability that subject i is observed at time t given $X_{i1},...,X_{iT}$, $V_{i1},...,V_{iT}$, $Y_{i1},...,Y_{iT}$; and $\pi_{it}(\hat{\lambda}) = p_{i1}(\hat{\lambda})\times\cdots\times p_{it}(\hat{\lambda})$. We also define the $T\times T$ diagonal matrix $\Delta_i(\lambda)$ as

$$\Delta_i(\lambda) = \begin{bmatrix} \frac{R_{i1}}{\pi_{i1}(\lambda)} & & & \text{O} \\ & \frac{R_{i2}}{\pi_{i2}(\lambda)} & & \\ & & \ddots & \\ \text{O} & & & \frac{R_{iT}}{\pi_{iT}(\lambda)} \end{bmatrix};$$

and similarly the matrix $\Delta_i(\hat{\lambda})$. If, as we are assuming here, all subjects are observed in the first wave, then the first element of $\Delta_i(\lambda)$ (and $\Delta_i(\hat{\lambda})$) is equal to 1 for all subjects.

In the WGEE methodology, instead of equations 2.4, we solve the following set of equations to get our estimate, $\hat{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$:

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \Delta_i(\hat{\lambda})(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{2.9}$$

Note that equations 2.9 differ for equations 2.4 only in the inclusion of the "weighting" matrix $\Delta_i(\hat{\lambda})$. This matrix has the effect of setting to zero any unobserved residual in the vector (of residuals) $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$, and weighting by the inverse of $\pi_{it}(\hat{\lambda})$ the corresponding observed residual in this vector.

Under 2.6, and provided the model for p_{it} is correctly specified, equation 2.9 has a root $\hat{\boldsymbol{\beta}}$ that is consistent for $\boldsymbol{\beta}$. Additionally, $\hat{\boldsymbol{\beta}}$ is unique with probability approaching to one and asymptotically normal, under mild regularity conditions. For the proofs see Robins et al. (1995).

The asymptotic variance of $\hat{\boldsymbol{\beta}}$ is given by

$$\mathrm{Var}_{\xi}(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \Gamma^{-1} C \left(\Gamma^{-1}\right)', \tag{2.10}$$

where

$$\Gamma = E_{\xi}\left\{ \frac{\partial}{\partial \boldsymbol{\beta}'} \left[\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \Delta_i(\lambda)(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \right] \right\},$$

$$C = I - J\Omega J',$$

$$I = E_{\xi}\left\{ \left[\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \Delta_i(\lambda)(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \right]^{\otimes 2} \right\},$$

$$J = E_{\xi}\left\{ \frac{\partial}{\partial \lambda'} \left[\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \Delta_i(\lambda)(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \right] \right\},$$

$$\Omega = \left[\mathrm{Var}\left\{ \frac{\partial l_i(\lambda)}{\partial \lambda} \right\} \right]^{-1},$$

and $A^{\otimes 2} = AA'$ for any matrix (or vector) A.

We will denote $\partial \boldsymbol{\mu}_i'/\partial \boldsymbol{\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ by $\partial \boldsymbol{\mu}_i'/\partial \hat{\boldsymbol{\beta}}$ for simplicity of notation. We can estimate the variance in 2.10 by

$$\hat{\mathrm{V}}\mathrm{ar}_{\xi}(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \hat{\Gamma}^{-1} \tilde{C} (\hat{\Gamma}^{-1})';$$

where

$$\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \Delta_i(\hat{\lambda}) \frac{\partial \boldsymbol{\mu}_i}{\partial \hat{\boldsymbol{\beta}}},$$

$$\tilde{C} = \hat{I} - \hat{J}\hat{\Omega}\hat{J}',$$

$$\hat{I} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \Delta_i(\hat{\lambda})(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i) \right]^{\otimes 2},$$

$$\hat{J} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \Lambda_i(\hat{\lambda}, \hat{\boldsymbol{\beta}}) \frac{\partial \pi_i(\lambda)}{\partial \hat{\lambda}'},$$

$$\Lambda_i(\hat{\lambda}, \hat{\boldsymbol{\beta}}) = \text{diag}\Big\{ \frac{R_{it}(y_{it} - \hat{\mu}_{it})}{[\pi_i(\hat{\lambda})]^2} \Big\},$$

$$\pi_i(\hat{\lambda}) = \big(\pi_{i1}(\hat{\lambda}), \ldots, \pi_{iT}(\hat{\lambda})\big)',$$

and $\hat{\Omega}$ is the inverse of the observed information from the partial pseudo likelihood 2.8.

Intermittent Missingness

The method in the previous subsection has the limitation of dealing only with (artificially) monotone datasets. Therefore, when the longitudinal dataset contains intermittent patterns, that method is obviously inefficient (does not use all data) and maybe even inconsistent if the reasons for dropping out of the study differ from the reasons for missing a wave (and coming back to the study). Robins et al. (1995) also propose an extension of WGEE applicable to longitudinal studies with intermittent missing observations, when this missing mechanism can be considered MAR.

We now present a detailed explanation of the WGEE methodology for intermittent longitudinal datasets. Again, we adopt the notation proposed by Hardin and Hilbe (2003).

We make the same assumptions about the model and the measurements as in the previous chapter, but here we additionally permit that some subjects who fail to respond at a given cycle may come back to the study at a later cycle. We intend to measure Y_i and X_i, as in equations 2.3, for all subjects, but due to dropouts and missing waves the full vector Y_i and matrix X_i are not always observed. We assume that in addition to Y_{it} and X_{it}, at time t we also observe a vector of covariates V_{it}, $t = 1, 2, ..., T$.

We define the response indicator variable $\tilde{R}_{it}$ in the following way,

$$\tilde{R}_{it} = \begin{cases} 1 & \text{if} \quad \text{subject } i \text{ is observed at time } t \\ 0 & \text{otherwise.} \end{cases}$$

And let $\tilde{Y}_{it} = \tilde{R}_{it} Y_{it}$, $\tilde{X}_{it} = \tilde{R}_{it} X_{it}$, and $\tilde{V}_{it} = \tilde{R}_{it} V_{it}$.

We assume that at any give time t, Y_{it}, X_{it}, and V_{it} are either all observed or all missing. In this subsection we allow dropouts and missing waves coming back; i.e. we allow the vector $\tilde{R}_i = (\tilde{R}_{i1}, \tilde{R}_{i2}, ..., \tilde{R}_{iT})'$ to take on any of 2^{T-1} possible realizations (i.e. any vector $r = (r_1, r_2, ..., r_T)'$ of zeros and ones of length T with first component equal to one). This method assumes that all subjects are observed at wave 1; i.e. $\tilde{R}_{i1} = 1$ for all subjects. And we redefine R_{it} in the following way,

$$R_{it} = \begin{cases} 1 & \text{if} \quad \tilde{R}_{i1} = \tilde{R}_{i2} = \cdots = \tilde{R}_{it} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

So, R_{it} is zero once a subject misses one wave.

In order to be able to extract some information from the subjects who return to the study, Robins et al. 1995 (and references therein), argue that one needs to assume the following equation, which is stronger than equation 2.6 (and stronger than MAR):

$$\begin{aligned} &P(\tilde{R}_{it}=1 \mid \tilde{R}_{i1},...,\tilde{R}_{i(t-1)},\ \tilde{X}_{i1},...,\tilde{X}_{i(t-1)},\ \tilde{V}_{i1},...,\tilde{V}_{i(t-1)},\ \tilde{Y}_{i1},...,\tilde{Y}_{i(t-1)},\\ &\qquad X_{i1},...,X_{iT},\ V_{i1},...,V_{iT},\ Y_{i1},...,Y_{iT})\\ &=P(\tilde{R}_{it}=1 \mid \tilde{R}_{i1},...,\tilde{R}_{i(t-1)},\ \tilde{X}_{i1},...,\tilde{X}_{i(t-1)},\ \tilde{V}_{i1},...,\tilde{V}_{i(t-1)},\ \tilde{Y}_{i1},...,\tilde{Y}_{i(t-1)})=p_{it} \end{aligned} \tag{2.11}$$

In words of Robins et al. (1995), this last assumption means that, the probability of being observed at time t, given the observed past, $\tilde{R}_{i1},...,\tilde{R}_{i(t-1)},\ \tilde{X}_{i1},...,\tilde{X}_{i(t-1)},\ \tilde{V}_{i1},...,\tilde{V}_{i(t-1)},\ \tilde{Y}_{i1},...,\tilde{Y}_{i(t-1)}$, through time $t-1$, does not depend on the unobserved past or present or on the future. We still assume that the probability p_{it} of being observed at time t, conditional on the observed past, $\tilde{R}_{i1},...,\tilde{R}_{i(t-1)},\ \tilde{X}_{i1},...,\tilde{X}_{i(t-1)},\ \tilde{V}_{i1},...,\tilde{V}_{i(t-1)},\ \tilde{Y}_{i1},...,\tilde{Y}_{i(t-1)}$, is bigger than zero for all times $t=2,...,T$. For an extension to the nonignorable case, see, for example, Rotnitzky et al. (1998).

We assume that the response probabilities p_{it} are a known function (taking values on $[0,1]$) of an unknown parameter λ, and the observed past, $\tilde{R}_{i1},...,\tilde{R}_{i(t-1)},\ \tilde{X}_{i1},...,\tilde{X}_{i(t-1)},\ \tilde{V}_{i1},...,\tilde{V}_{i(t-1)},\ \tilde{Y}_{i1},...,\tilde{Y}_{i(t-1)}$; i.e.

$$p_{it}=p_{it}(\tilde{R}_{i1},...,\tilde{R}_{i(t-1)},\ \tilde{X}_{i1},...,\tilde{X}_{i(t-1)},\ \tilde{V}_{i1},...,\tilde{V}_{i(t-1)},\ \tilde{Y}_{i1},...,\tilde{Y}_{i(t-1)};\ \lambda).$$

We will assume that p_{it} follows the logistic regression model

$$\begin{aligned} \text{logit}(p_{it}) &= \log\left(\frac{p_{it}}{1-p_{it}}\right)\\ &= \lambda' h(\tilde{R}_{i1},...,\tilde{R}_{i(t-1)},\ \tilde{X}_{i1},...,\tilde{X}_{i(t-1)},\ \tilde{V}_{i1},...,\tilde{V}_{i(t-1)},\ \tilde{Y}_{i1},...,\tilde{Y}_{i(t-1)}) \end{aligned} \tag{2.12}$$

where $\lambda'=(\lambda_1,...,\lambda_q)$, and $h(\cdot)$ is a $q\times 1$ known vector function of $\tilde{R}_{i1},...,\tilde{R}_{i(t-1)},\ \tilde{X}_{i1},...,\tilde{X}_{i(t-1)},\ \tilde{V}_{i1},...,\tilde{V}_{i(t-1)},\ \tilde{Y}_{i1},...,\tilde{Y}_{i(t-1)}$.

Now we let $\hat{\lambda}$ solve the estimating equation $\sum_{i=1}^{n} S_{\lambda,i}(\lambda)=0$, where

$$S_{\lambda,i}(\lambda)=\frac{\partial}{\partial\lambda}\log\left\{\prod_{t=2}^{T}[p_{it}(\lambda)]^{\tilde{R}_{it}}[1-p_{it}(\lambda)]^{1-\tilde{R}_{it}}\right\}=\sum_{t=2}^{T}[\tilde{R}_{it}-p_{it}(\lambda)]\frac{\partial \text{logit}(p_{it})}{\partial\lambda}.$$

We define, for a subject i with nonresponse history $\tilde{R}_i=r$,

$$\begin{aligned} &\pi_i(r,\lambda)\\ &=\prod_{t=2}^{T}\left\{p_{it}(r_1,...,r_{t-1},\ r_1X_{i1},...,r_{t-1}X_{i(t-1)},\ r_1V_{i1},...,r_{t-1}V_{i(t-1)},\ r_1Y_{i1},...,r_{t-1}Y_{i(t-1)};\lambda)^{r_t}\times\right.\\ &\quad\left.[1-p_{it}(r_1,...,r_{t-1},\ r_1X_{i1},...,r_{t-1}X_{i(t-1)},\ r_1V_{i1},...,r_{t-1}V_{i(t-1)},\ r_1Y_{i1},...,r_{t-1}Y_{i(t-1)};\lambda)]^{1-r_t}\right\}; \end{aligned}$$

which, under 2.11, is equal to

$$\pi_i(r,\lambda)=P(\tilde{R}_i=r\,\big|\,X_{i1},...,X_{iT},\ V_{i1},...,V_{iT},\ Y_{i1},...,Y_{iT});$$

that is, it is the probability that a subject i with given $X_{i1},...,X_{iT}$, $V_{i1},...,V_{iT}$, $Y_{i1},...,Y_{iT}$ has a nonresponse pattern given by the vector r.

We now let $\Phi_i = \{\Phi_{i,r} : r \neq (1,1,...,1)\}$, where $\Phi_{i,r}$ is, for each $r \neq (1,1,...,1)$, a known ν-dimensional function of $r_1X_{i1},...,r_TX_{iT}$, $r_1V_{i1},...,r_TV_{iT}$, $r_1Y_{i1},...,r_TY_{iT}$ selected by the investigator (Robins et al., 1995). We define

$$A_i(\Phi,\hat{\lambda}) = \frac{R_{iT}}{\pi_{iT}(\hat{\lambda})}\sum\nolimits_{r\neq(1,1,...,1)}\pi_i(r,\hat{\lambda})\Phi_{i,r} - \sum\nolimits_{r\neq(1,1,...,1)} I(\tilde{R}_i = r)\Phi_{i,r},$$

where

$$I(\tilde{R}_i = r) = \begin{cases} 1 & \text{if} \quad \tilde{R}_i = r \\ 0 & \text{otherwise.} \end{cases}$$

Robins et al. (1995) show that under 2.11, $A_i(\Phi,\hat{\lambda})$ has mean zero for all subjects i.

In this case, instead of equations 2.9, we solve iteratively the following set of equations to get our estimate, $\tilde{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$:

$$\sum_{i=1}^{n}\left\{\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\Delta_i(\hat{\lambda})(\boldsymbol{y}_i-\boldsymbol{\mu}_i) - \hat{\theta}A_i(\Phi,\hat{\lambda})\right\} = \mathbf{0}; \tag{2.13}$$

where $\Delta_i(\hat{\lambda})$ is defined as in the previous subsection but replacing R_{it} by the one in this subsection; let $U_i(\boldsymbol{\beta},\hat{\lambda}) = {}^{\partial\boldsymbol{\mu}_i'}/_{\partial\boldsymbol{\beta}}V_i^{-1}\Delta_i(\hat{\lambda})(\boldsymbol{y}_i-\boldsymbol{\mu}_i)$, then $\hat{\theta} = \hat{\theta}_1\hat{\theta}_2^{-1}$,

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} e\big(U_i(\boldsymbol{\beta},\hat{\lambda}),\, S_{\lambda,i}(\hat{\lambda})\big)e\big(A_i(\Phi,\hat{\lambda}),\, S_{\lambda,i}(\hat{\lambda})\big)',$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} e\big(A_i(\Phi,\hat{\lambda}),\, S_{\lambda,i}(\hat{\lambda})\big)e\big(A_i(\Phi,\hat{\lambda}),\, S_{\lambda,i}(\hat{\lambda})\big)',$$

where

$$e(P_i,Q_i) = P_i - \big(\sum_{i=1}^{n}P_iQ_i'\big)\big(\sum_{i=1}^{n}Q_iQ_i'\big)^{-1}Q_i'$$

is the residual for subject i from the multivariate regression of the vectors P_i on the vectors Q_i, $i = 1,2,...,n$. Note that $U_i(\boldsymbol{\beta},\hat{\lambda})$ is $p\times 1$, $S_{\lambda,i}(\hat{\lambda})$ is $q\times 1$, and $A_i(\Phi,\hat{\lambda})$ is $\nu\times 1$.

Under 2.11, and provided the model for p_{it} is correctly specified, equation 2.13 has a root $\tilde{\boldsymbol{\beta}}$ that is consistent for $\boldsymbol{\beta}$. Additionally, $\tilde{\boldsymbol{\beta}}$ is unique with probability approaching to one and asymptotically normal, under mild regularity conditions; for the proofs see Robins et al. (1995). They also argue that increasing the dimension, ν, of $\Phi_{i,r}$ never increases the asymptotic variance of $\tilde{\boldsymbol{\beta}}$ and usually decreases it.

The asymptotic variance of $\tilde{\boldsymbol{\beta}}$ is given by

$$\mathrm{Var}_\xi(\tilde{\boldsymbol{\beta}}) = \frac{1}{n}\Gamma^{-1}\mathrm{Var}[e(U_i(\boldsymbol{\beta},\hat{\lambda}),B_i)]\left(\Gamma^{-1}\right)', \tag{2.14}$$

$$\Gamma = E_\xi\left\{\frac{\partial}{\partial\boldsymbol{\beta}'}\left[\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\Delta_i(\lambda)(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\right]\right\},$$

$$B_i' = \left(A_i(\Phi,\lambda)',\, S_{\lambda,i}(\lambda)'\right).$$

Robins et al. (1995) claim that the estimator obtained by applying the method in the previous subsection to the artificially monotone dataset (i.e. ignoring any subject's data after a missing wave) is never more efficient than $\tilde{\boldsymbol{\beta}}$.

We can estimate the variance of $\tilde{\boldsymbol{\beta}}$ in 2.14 as

$$\hat{\mathrm{V}}\mathrm{ar}_{\xi}(\tilde{\boldsymbol{\beta}}) = \frac{1}{n^2}\hat{\Gamma}^{-1}\sum_{i=1}^{n}[e(U_i(\tilde{\boldsymbol{\beta}},\hat{\lambda}),\hat{B}_i)]^{\otimes 2}(\hat{\Gamma}^{-1})';$$

where

$$\hat{B}_i' = \left(A_i(\Phi,\hat{\lambda})',\, S_{\lambda,i}(\hat{\lambda})'\right)$$

and

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{\mu}_i'}{\partial\tilde{\boldsymbol{\beta}}}\hat{V}_i^{-1}\Delta_i(\hat{\lambda})\frac{\partial\boldsymbol{\mu}_i}{\partial\tilde{\boldsymbol{\beta}}}.$$

2.2.4 Other Methods

Besides the re-weighting method of section 2.2.3, there are some other alternatives to marginal estimation in longitudinal studies with missing responses, beyond the "naïve" complete case or available case analyses.

A common method of handling nonresponse, especially in pharmaceutical studies, is the so-called "last observation carried forward" (LOCF); some examples are given by Ali and Talukder (2005). The LOCF imputation method substitutes any missing value in a variable by the last available value of that variable (at a previous wave); the filled-in dataset is then analysed by usual complete case methods. This technique rests on the strong assumption that the unobserved values remain unchanged after dropout. Liu and Gould (2002) and Shao and Zhong (2006) show some examples in which the LOCF method produces biased estimates of treatment effects, and leads to underestimation of variability. Cook et al. (2004) additionally describe some instances where it causes biases in estimation of regression coefficients and inflated type I error rates. They argue that the degree to which estimation from LOCF is valid or invalid depends on the MCAR vs. MAR mechanism and on whether there is a trend in the variable of interest over time.

Another alternative for handling missing values in longitudinal studies is multiple imputation (MI). Here, the missing observations are imputed, or filled-in, by some "proper" mechanism, to obtain a completed dataset. This process is repeated a certain number of times to get several completed datasets. Each of them is analyzed using standard complete case methods, and the results from the different datasets are then combined to obtain the point and variance estimators. For a comprehensive treatment of MI (including the definition of "proper") see in Rubin (1987). Some examples of application of this method are presented in Lavori et al. (1995), Liu and Gould (2002), and Taylor et al. (2002).

For complex survey data, Fay (1996) (and other references therein) shows some cases where the MI technique leads to inconsistent variance estimation; for example "in the common situation of imputation cutting across sample clusters" (Rao and Shao, 1992). Binder and Sun

(1996) show the conditions necessary for the imputation to be proper in a complex survey situation, and argue that satisfying these conditions is difficult. The fact that the imputations are improper results in strong overestimation of variability for some domain estimates (when the domain indicator is not included in the imputation model); this problem is described in Fay (1996) and the comments to that paper. Since domain estimation is extensively used in surveys and "many domain estimates will not have been identified at the time that the imputation was carried out... MI is not generally recommended for public use data files" (Kim et al., 2006). For ignorable sampling mechanisms, Kim et al. (2006) also give expressions for the bias of the MI variance estimator, and propose an adjustment by incorporating the survey weights in the imputation model. Additionally, according to Rao and Shao (1992), "several statistical agencies seem to prefer single imputation, mainly due to operational difficulties in maintaining multiple complete data sets, especially in large-scale surveys."

Likelihood methods also exist for estimation with missing values in longitudinal studies. Some examples include Baker (1995), Fitzmaurice et al. (1996), Galecki et al. (2001), and Yi and Thompson (2005).

Chapter 3

The Pseudo-GEE Approach to the Analysis of Longitudinal Surveys with Complete Data

3.1 The Joint Randomization Framework

The GEE method needs to be modified to be applied to data obtained from complex surveys. The sampled units cannot be regarded as independent (clustering); and also, they all should not account for the same amount in the estimation procedure since they are likely to have been selected with varying probabilities. We begin with a brief discussion about different approaches to inference from complex surveys.

There are three popular ways of inference from surveys of finite populations. In the pure "model-based" approach the parameters of interest are parameters in a statistical model (superpopulation model). Under this setting the design characteristics are ignored, sampled individuals are treated as i.i.d. observations, and all the inferences are carried out and evaluated only with respect to the model. If the model being fitted is correct, one should use optimal estimators with respect to the model, which usually means, ignoring the sample characteristics (Binder and Roberts, 2003). They also show some examples in which this kind of estimators may even be better (smaller design-based MSE) than design-based counterparts for estimating finite population parameters. Scott and Smith (1974) argue that when strong knowledge about the model exists, the sampling mechanism is irrelevant for estimation. On the other hand, if the assumed model is misspecified, the results obtained by model-based methods can be invalid; in the sense that point estimators may be inconsistent for the superpopulation parameters and/or that variance estimators may be incorrect.

A commonly used mode of inference with samples from finite populations is known as "design-based" approach. Here the parameters of interest are finite population quantities, "regarded as descriptive parameters to be estimated" (Binder and Roberts, 2003); all the variables of interest are treated as nonrandom quantities; and the procedures are evaluated only with respect to the properties of the selection mechanism of the sample. The biggest advantage of this kind of inference is that the estimators are usually design unbiased or consistent for the finite

population values, regardless of any model assumptions. Even more, "design-based estimates tend to give valid inferences, even in some cases when the model is misspecified," compared to pure model-based estimators (Binder and Roberts, 2003). Obviously, design-based estimators are less efficient than model-based ones when the model *is* true; but "for very large samples, this loss may not be too serious, given the extra robustness achieved from the design-based approach."

The other popular way of inference is the "model-assisted" approach. Here, again, interest lies exclusively on finite population parameters, all the observed quantities are regarded as nonrandom, and all procedures are judged only with respect to the sampling design used. The difference between this method and the design-based is that, in the present, superpopulation models are borrowed to motivate approaches, specifically estimators. This last method has the advantage over the design-based one that if the model used to motivate the estimators is justifiable (i.e. is a good representation of the finite population) it increases efficiency.

A totally different method of inference from surveys is sometimes referred to as "joint randomization" inference. In this case one is interested in superpopulation models (and possibly causal relationships) thought to have generated the finite population, from which the sample is obtained to make the inferences. The sample can be thought of as a second phase of sampling from the superpopulation (Binder and Roberts, 2003). The main difference of this case and the previous ones is that here both parts of the process, the randomization imposed by the model generating the finite population values and the randomization introduced by the sampling selection, are taken into account for the inferences.

Although this strategy is not used nearly as often as the previous three, it comprises some characteristics that make it appealing for analytic uses of survey data. For one thing, we should evaluate the model and evaluate the estimators with respect this model because "the ultimate objective of causal modelling may be to develop the widely applicable models assumed with the model-based approach" (Kalton, 1983). For another, when drawing inferences from survey data, it is usually not appropriate to ignore the design features, so that plain model-based techniques are generally not suited for survey data. The design features of the sample should not, in general, be ignored, even under model inferences, mainly due to lack of independence among the sampled units because of clustering in the population (and in the sample); and also because of differential selection probabilities, which make the elements not "identical". Furthermore, as noted by Kalton (1983), pure model-based methods can be severely affected by things like excluding important variables or interaction terms. Whereas in such a situation, the inclusion of the design characteristics yields "the best fit of that model for the surveyed population, and hence also a good fit for similar populations where 'similar' relates to the excluded variables." Another reason for a joint randomization approach is that, even under the design-based approach, certain optimality criteria necessarily rely on models (as in Wu, 2003). And finally, sometimes it is the only appropriate method of inference because of the way in which the data are collected (as in Chen et al., 2004).

In this thesis we will follow this approach because what we are really interested in is model parameters; i.e. we are interested in the effect of some covariates on an outcome variable, or how the outcome variable changes with changes in the covariates. And the survey we will use for our analyses is complex in the sense that the units are sampled in clusters (more specifically

in multiple stages) and selected with differential probabilities and thus have different weights. Additionally, we will be dealing with missing responses; and then it is necessary to posit a model, either explicit or implicit. It is reasonable to use the "best" possible model, which should be the main model, for the response mechanism. Therefore, it is natural to include, in the evaluation of the estimators, the random mechanism imposed by the main model.

In the next section we first describe the pseudo-GEE approach to the analysis of longitudinal survey data with complete responses. The use of survey weights under the estimating equation approach has been examined by several authors, including Godambe and Thompson (1986), Binder and Patak (1994) and Godambe (1995), among others. The consistency of the resulting estimators, however, has not been formally established in these earlier investigations. Several authors have investigated the asymptotic properties of estimating equations under an assumed GEE model, for example Inagaki (1973), Yuan and Jennrich (1998), and Shao (2003), but not in survey settings. We show that the pseudo-GEE estimators are consistent under the proposed joint randomization framework.

A similar framework has been used by Rubin-Bleuer and Schiopu Kratina (2005) under a more rigorous treatment using a product probability space. We take a more pragmatic approach in this thesis; depending on the circumstance, we use a conditional argument with a particular order of the involved randomizations. We are able to do this under the assumption that the involved randomizations from different sources are unconfounded[1]; we can use the order we see fit in each case.

For asymptotic development, we assume that there is a sequence of finite populations, indexed by ν. Both, the population size N_ν and the sample size n_ν depend on ν. All limiting processes are understood as $\nu \to \infty$. We assume that $N_\nu \to \infty$ and $n_\nu \to \infty$ as $\nu \to \infty$. Nonetheless, for simplicity, we will drop the dependence on ν, from the notation of N_ν and n_ν, and use $N \to \infty$ or $n \to \infty$, instead.

3.2 The Pseudo-GEE with Complete Data

We assume that we have a GEE model (satisfying the four items described in section 2.1.2) about which we want to make inferences (i.e. inferences about the $\boldsymbol{\beta}$ coefficients); we denote this model by ξ. In other words, we think of an infinite superpopulation satisfying the model

$$\xi: \begin{cases} E[Y_{ij}|X_{ij}] = \mu_{ij} = g^{-1}(\eta_{ij}) = g^{-1}(X_{ij}'\boldsymbol{\beta});\ i = 1,2,...;\ j = 1,2,... \\ \text{Var}[Y_{ij}|X_{ij}] = \phi\, v(\mu_{ij});\ i = 1,2,...;\ j = 1,2,... \\ \text{Cov}[Y_i|X_i] = A_i^{1/2}\mathbf{R}_i(\alpha)A_i^{1/2};\ A_i = \text{diag}[\phi\, v(\mu_{ij})];\ \mathbf{R}(\alpha) = \text{working correlation matrix} \\ Y_k \text{ and } Y_l \text{ are independent vectors conditional on } X_k \text{ and } X_l,\ \forall\, k \neq l; \end{cases}$$

with the requirements specified in section 2.1.2. For notational simplicity, from now on we will drop the dependence of the (conditional) mean and variance of Y_{ij} on X_{ij}. So that $E[Y_{ij}]$ and $\text{Var}[Y_{ij}]$ denote the expected value and variance of Y_{ij}, conditional on the set of covariates X_{ij}.

[1] See section 4.2 (page 43) for more detail about unconfoundedness.

Now, we assume that the finite population is a random sample of N elements drawn from model ξ. Assume for the moment that we are able to "sample" the whole finite population, i.e. that we could have a census of all the N elements. Since this census is in fact a random sample from model ξ, by the theory in section 2.1.2, we could solve the following set of estimating equations to obtain B:

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{3.1}$$

B is the so-called "census estimator" of $\boldsymbol{\beta}$ if we were able to sample the whole N elements. Note that B is also denoted by $\boldsymbol{\beta}_N$ or $\hat{\boldsymbol{\beta}}_N$ indistinguishably in the literature. This estimator plays an important role in our theoretical development but has no practical value. The real question of interest is how to make inference about $\boldsymbol{\beta}$ based on a survey sample selected from the finite population.

We denote such a sample (of size n) by s and the sampling mechanism by π, which will usually be complex. We assume that each subject i has associated with it a "survey weight" w_i. This weight will, basically, be the inverse of the probability of selection of unit i, but will also be adjusted to account for things like nonresponse and calibration to known finite population totals. An estimate of the finite population size, N, is

$$\hat{N} = \sum_s w_i \, .$$

If we treat the left-hand side of equation 3.1 as a finite population total, we can estimate it based on sample s, using the well-known Horvitz-Thompson estimator (1952):

$$\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i),$$

where the sum is over the n elements in s. The sample based estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_n$, is defined as the solution to the following set of estimating equations:

$$\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{3.2}$$

The Newton-Raphson procedure described in section 2.1.2 can be modified to obtain a solution to 3.2. The updating step now becomes

$$\boldsymbol{\beta}_{(l+1)} = \boldsymbol{\beta}_{(l)} + \left[\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_{(l)}} V_{i(l)}^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}_{(l)}}\right]^{-1} \left[\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_{(l)}} V_{i(l)}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_{i(l)})\right].$$

With these modifications, the dispersion parameter ϕ is estimated by

$$\hat{\phi} = \frac{\sum_{i\in s} w_i \sum_{j=1}^{T_i} e_{ij}^2}{(\sum_s w_i T_i) - p} = \frac{\sum_{i\in s} w_i \sum_{j=1}^{T_i} (y_{ij} - \hat{\mu}_{ij})^2 / v(\hat{\mu}_{ij})}{(\sum_s w_i T_i) - p}; \tag{3.3}$$

and if the within-subject association is unspecified (and the same for all subjects), we estimate α by

$$\hat{\alpha}_{jk} = \frac{\sum_{i\in s} w_i e_{ij} e_{ik}}{[(\sum_s w_i) - p]\hat{\phi}} = \frac{\sum_{i\in s} w_i (y_{ij} - \hat{\mu}_{ij})(y_{ik} - \hat{\mu}_{ik}) / \sqrt{v(\hat{\mu}_{ij}) v(\hat{\mu}_{ik})}}{[(\sum_s w_i) - p]\hat{\phi}}. \tag{3.4}$$

and the standardized residuals, e_{ij}, are given by the same form as in section 2.1.2.

Because of these changes, from n to $\hat{N}$ and from $\sum_{i=1}^{n} T_i$ to $\sum_s w_i T_i$, usual GEE software procedures like `gee` in R or `genmod` in SAS are not recommended for survey data. Even if one specifies the `weight` variable as the survey weights w_i, these procedures do not carry out the appropriate modification of $\hat{\phi}$ and $\hat{\alpha}$.

With respect to estimation of uncertainty about $\hat{\boldsymbol{\beta}}_n$, we should have in mind that we are following joint randomization inference. Therefore we ought to calculate the variance of it with respect to both randomization mechanisms: the randomization induced by the model *and* the randomization induced by the sampling scheme. As Kovacevic and Rai (2002) point out, "in general the total variance should be used for inference about the superpopulation parameters because it accounts for both variabilities." We use subscripts ξ or π to indicate under the model or under the design respectively; and use the double subscript $\xi\pi$ to indicate under the mixed randomization.

The estimator $\hat{\boldsymbol{\beta}}_n$ is consistent for $\boldsymbol{\beta}$ jointly under the model and the design, as the following theorem states:

Theorem 3.1. *Let $s_n(\boldsymbol{\beta}) = \sum_s w_i \psi_i(Y_i, \boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \Theta \subset \mathbb{R}^p$ and $\psi_i(Y_i, \boldsymbol{\beta})$ is a function from $\mathbb{R}^{T_i} \times \Theta$ to $\mathbb{R}^p$; let $\boldsymbol{\beta}_0 \in \Theta$ be such that $E_{\xi\pi}[s_n(\boldsymbol{\beta}_0)] = 0$; let $h_i(Y_i) = \sup_{\boldsymbol{\beta}\in\Theta} \|\psi_i(Y_i, \boldsymbol{\beta})\|$, $i = 1, 2, \ldots$, where $\|\cdot\|$ is the usual $\mathscr{L}_1$ norm. Suppose that*

1. *$\sup_i E_\xi[h_i^2(Y_i)] < \infty$ and $\sup_i E_\xi ||Y_i|| < \infty$;*
2. *For any $c > 0$ and sequence $\{y_i\}$ satisfying $\|y_i\| \le c$, the sequence of functions $\{g_i(\boldsymbol{\beta}) = \psi_i(y_i, \boldsymbol{\beta})\}$ is equicontinuous on any open subset of Θ;*
3. *The function $\Delta_N(\boldsymbol{\beta}) = E_{\xi\pi}[N^{-1} s_n(\boldsymbol{\beta})]$ has the property that $\inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\varepsilon} |\Delta_N(\boldsymbol{\beta})| > 0$ for any $\varepsilon > 0$;*
4. *There exists a $\hat{\boldsymbol{\beta}}_n \in \Theta$ that is solution to $s_n(\boldsymbol{\beta}) = 0$, i.e. $\hat{\boldsymbol{\beta}}_n$ is the pseudo-GEE estimator of $\boldsymbol{\beta}$ such that $s_n(\hat{\boldsymbol{\beta}}_n) = 0$;*
5. *$\hat{\boldsymbol{\beta}}_n = O_p(1)$;*
6. *The design weights w_i satisfy $N^{-1}\sum_s w_i Z_i - N^{-1}\sum_{i=1}^{N} Z_i = O_p(1/\sqrt{n})$ for any variable Z such that $N^{-1}\sum_{i=1}^{N} Z_i^2 = O(1)$;*

then $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$, where "$p$" denotes in probability w.r.t. both the model ξ and the sampling design π.

Condition 5 is weaker than assuming the parameter space is compact, which is what Robins et al. (1995) assumed for their results. Here the "p" in $O_p(1)$ means in probability with respect to the joint $\xi\pi$ distribution.

In condition 6 the "p" in $O_p(1/\sqrt{n})$ means under the distribution induced by the design π. This condition is weaker than assuming $N^{-1}\sum_s w_i Z_i$ is asymptotically normally distributed. That is, if $\widehat{\bar{Z}}_{HT} = N^{-1}\sum_s w_i Z_i \dot{\sim} N(\bar{Z}, \sigma^2/n)$, then condition 6 is satisfied. "Hájek (1960, 1964) established the asymptotic normality of $\widehat{\bar{Y}}_{HT}$ under simple random sampling and rejective sampling with unequal selection probabilities. Víšek (1979) established the asymptotic normality of $\widehat{\bar{Y}}_{HT}$ for the well-known Rao-Sampford method of unequal probability sampling without replacement" (Wu and Rao, 2006).

The joint variance of $\hat{\boldsymbol{\beta}}_n$ is given by

$$\text{Var}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n) = \text{Var}_{\xi}[E_{\pi}(\hat{\boldsymbol{\beta}}_n)] + E_{\xi}[\text{Var}_{\pi}(\hat{\boldsymbol{\beta}}_n)]. \tag{3.5}$$

The first component in 3.5, $\text{Var}_{\xi}[E_{\pi}(\hat{\boldsymbol{\beta}}_n)]$, is called the "model variance component" and represents the variance in a census fit to the model, using data from the entire finite population. The second component, $E_{\xi}[\text{Var}_{\pi}(\hat{\boldsymbol{\beta}}_n)]$, is called the "design variance component" or "sampling variance component" and represents the additional variance contributed by sampling of the finite population; it comes from the fact that a sample of n elements is observed rather than the entire finite population of N elements (Särndal et al., 1992).

Let $B^{\diamond} = E_{\pi}(\hat{\boldsymbol{\beta}}_n)$; i.e. $B^{\diamond}$ is the conceptual finite population quantity which is unbiasedly estimated by $\hat{\boldsymbol{\beta}}_n$. If $\text{Var}_{\xi}[B^{\diamond}]$ has the usual order of $1/N$, and suppose that the sampling fraction n/N is small or negligible, then the leading term in the joint variance is $E_{\xi}[\text{Var}_{\pi}(\hat{\boldsymbol{\beta}}_n)]$. Note that the $B^{\diamond}$ defined above is usually not identical to B, the solution to 3.1. Therefore we can write

$$\text{Var}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n) \doteq E_{\xi}[\text{Var}_{\pi}(\hat{\boldsymbol{\beta}}_n)]. \tag{3.6}$$

We can estimate the joint variance of $\hat{\boldsymbol{\beta}}_n$ in 3.6 with $\hat{\text{V}}\text{ar}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n) = \hat{\text{V}}\text{ar}_{\pi}(\hat{\boldsymbol{\beta}}_n)$; where $\hat{\text{V}}\text{ar}_{\pi}(\hat{\boldsymbol{\beta}}_n)$ is an estimator of the design variance of $\hat{\boldsymbol{\beta}}_n$, which can be constructed using Taylor Linearization (as in Binder, 1983) or replication techniques. The estimator $\hat{\text{V}}\text{ar}_{\xi\pi}(\hat{\boldsymbol{\beta}}_n)$ is approximately unbiased under the joint randomization since $E_{\xi\pi}[\hat{\text{V}}\text{ar}_{\pi}(\hat{\boldsymbol{\beta}}_n)] = E_{\xi}\{E_{\pi}[\hat{\text{V}}\text{ar}_{\pi}(\hat{\boldsymbol{\beta}}_n)]\} \doteq E_{\xi}[\text{Var}_{\pi}(\hat{\boldsymbol{\beta}}_n)]$.

3.3 Proof of Consistency of the Pseudo-GEE Estimators

Theorem 3.1 establishes the weak consistency of the pseudo-GEE estimator $\hat{\boldsymbol{\beta}}_n$ under the joint randomization of both the model and the sampling design. The following lemma, adapted from Lemma 5.3. of Shao (2003), plays a key role in proving our theorem.

Lemma 3.2. *Suppose that*

1. Θ *is a compact subset of* $\mathbb{R}^p$;
2. $\sup_i E_{\xi}[h_i^2(Y_i)] < \infty$ *and* $\sup_i E_{\xi}||Y_i|| < \infty$, *where* $h_i(Y_i) = \sup_{\boldsymbol{\beta}\in\Theta} \|\psi_i(Y_i, \boldsymbol{\beta})\|$ *and* $\psi_i(Y_i, \boldsymbol{\beta})$ *is a function from* $\mathbb{R}^{T_i} \times \Theta$ *to* $\mathbb{R}^p$;

3. *For any $c > 0$ and sequence $\{y_i\}$ satisfying $\|y_i\| \leq c$, the sequence of functions $\{g_i(\boldsymbol{\beta}) = \psi_i(y_i, \boldsymbol{\beta})\}$ is equicontinuous on any open subset of Θ;*

4. *$N^{-1}\sum_s w_i Z_i - N^{-1}\sum_{i=1}^N Z_i = O_p(1/\sqrt{n})$, where Z is any variable independent of the design, for which $N^{-1}\sum_{i=1}^N Z_i^2 = O(1)$;*

then, as $n \to \infty$ (and $N \to \infty$),

$$\sup_{\boldsymbol{\beta}\in\Theta} \left|\left| \frac{1}{N} s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta}) \right|\right| \xrightarrow{p} 0,$$

where $s_n(\boldsymbol{\beta}) = \sum_s w_i \psi_i(Y_i, \boldsymbol{\beta})$ and $\Delta_N(\boldsymbol{\beta}) = E_{\xi\pi}[N^{-1} s_n(\boldsymbol{\beta})] = N^{-1}\sum_{i=1}^N E_\xi[\psi_i(Y_i, \boldsymbol{\beta})]$.

Proof of Lemma 3.2.
Without loss of generality we assume that the ψ_i's are scalar; if each of the components (in absolute value) of a vector goes to zero in probability, then the norm of the vector goes to zero in probability. By Hölder's inequality (with $p = q = 2$) and Markov's inequality, for any $c > 0$,

$$\begin{aligned} E_\xi\left[h_i(Y_i) I(\|Y_i\| > c)\right] &\leq \left[E_\xi[h_i^2(Y_i)]\right]^{1/2} \left[E_\xi I(\|Y_i\| > c)\right]^{1/2} \\ &\leq \left[\sup_i E_\xi[h_i^2(Y_i)]\right]^{1/2} \left[\frac{\sup_i E_\xi\|Y_i\|}{c}\right]^{1/2}. \end{aligned}$$

Let $c_0 = \sup_i E_\xi[h_i^2(Y_i)]$ and $c_1 = \sup_i E_\xi||Y_i||$ (both $< \infty$ by assumption 2 in the lemma). Then, **for all** i,

$$E_\xi\left[h_i(Y_i) I(\|Y_i\| > c)\right] \leq c_0^{1/2} c_1^{1/2} c^{-1/2} = O(c^{-1/2}). \tag{3.7}$$

For any $\varepsilon > 0$ and $\mathscr{O} \subset \Theta$, Markov's inequality and 3.7 imply that:

$$\begin{aligned} P_{\xi\pi}\Big(\frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) I(\|Y_i\| > c) > \varepsilon\Big) &\leq P_{\xi\pi}\Big(\frac{1}{N}\sum_s w_i h_i(Y_i) I(\|Y_i\| > c) > \varepsilon\Big) \\ &\leq \frac{1}{\varepsilon} E_{\xi\pi}\Big[\frac{1}{N}\sum_s w_i h_i(Y_i) I(\|Y_i\| > c)\Big] \\ &\leq \frac{1}{\varepsilon}\sup_i E_\xi\left[h_i(Y_i) I(\|Y_i\| > c)\right] \\ &= O(c^{-1/2}); \end{aligned}$$

based on this, and considering that

$$-\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) = \sup_{\boldsymbol{\beta}\in\mathscr{O}} \{-\psi_i(Y_i, \boldsymbol{\beta})\} \leq \sup_{\boldsymbol{\beta}\in\mathscr{O}} ||\psi_i(Y_i, \boldsymbol{\beta})||,$$

we have that for any $\mathscr{O} \subset \Theta$, as $c \to \infty$,

$$P_{\xi\pi}\left(\frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) I(\|Y_i\| > c) - \frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) I(\|Y_i\| > c) > \varepsilon\right) \to 0. \tag{3.8}$$

Now, by the equicontinuity of $\{g_i(\boldsymbol{\beta}) = \psi_i(Y_i, \boldsymbol{\beta})\}$ over i and Θ, given any $\varepsilon > 0$, there exists a δ, such that for any open ball $\mathscr{O}$ with radius less than δ, if $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathscr{O}$, we have:

$$\left|\psi_i(Y, \boldsymbol{\beta}_1) I(\|Y\| \leq c) - \psi_i(Y, \boldsymbol{\beta}_2) I(\|Y\| \leq c)\right| \leq \varepsilon/4,$$

for any Y. This implies that

$$\sup_{\boldsymbol{\beta}_1 \in \mathscr{O}} \sup_{\boldsymbol{\beta}_2 \in \mathscr{O}} [\psi_i(Y, \boldsymbol{\beta}_1) I(\|Y\| \leq c) - \psi_i(Y, \boldsymbol{\beta}_2) I(\|Y\| \leq c)] \leq \varepsilon/4,$$

and therefore,

$$\frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) I(\|Y_i\| \leq c) - \frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) I(\|Y_i\| \leq c) \leq \left[\frac{1}{N}\sum_s w_i\right] \frac{\varepsilon}{4}; \quad (3.9)$$

note that (by assumption 4) $N^{-1}\sum_s w_i = 1 + O_p(1/\sqrt{n})$, then we have $P_\pi(N^{-1}\sum_s w_i > 2) \to 0$; consequently, using 3.8 and 3.9, as $n \to \infty$,

$$P_{\xi\pi}\left(\frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) - \frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) > \frac{\varepsilon}{2}\right) \to 0. \quad (3.10)$$

Now, for the same $\mathscr{O}$, by the WLLN for independent random variables, and condition 2, we know that

$$P_\xi\left(\left\|\frac{1}{N}\sum_{i=1}^N \inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^N E_\xi\left[\inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta})\right]\right\| > \frac{\varepsilon}{2}\right) \to 0. \quad (3.11)$$

Besides, note that $|\inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta})| \leq h_i(Y_i)$; so that by assumption 2 we can apply assumption 4 to $\inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta})$, to get

$$\left\|\frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^N \inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta})\right\| = O_p(n^{-1/2}) = o_p(1). \quad (3.12)$$

It follows from 3.11 and 3.12 that

$$P_{\xi\pi}\left(\left\|\frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^N E_\xi\left[\inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta})\right]\right\| > \varepsilon\right) \to 0. \quad (3.13)$$

Hence, by 3.10 and 3.13,

$$P_{\xi\pi}\left(\frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^N E_\xi\left[\inf_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta})\right] > \varepsilon\right) \to 0. \quad (3.14)$$

Now let $H_n(\boldsymbol{\beta}) = N^{-1}\sum_s w_i \psi_i(Y_i, \boldsymbol{\beta}) - N^{-1}\sum_{i=1}^N E_\xi[\psi_i(Y_i, \boldsymbol{\beta})]$; then

$$\sup_{\boldsymbol{\beta} \in \mathscr{O}} H_n(\boldsymbol{\beta}) \leq \frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta} \in \mathscr{O}} \psi_i(Y_i, \boldsymbol{\beta}) + \frac{1}{N}\sum_{i=1}^N \sup_{\boldsymbol{\beta} \in \mathscr{O}} E_\xi[-\psi_i(Y_i, \boldsymbol{\beta})]$$

$$\begin{aligned} &\leq \frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta}) + \frac{1}{N}\sum_{i=1}^{N} E_\xi\Big[\sup_{\boldsymbol{\beta}\in\mathscr{O}}\{-\psi_i(Y_i,\boldsymbol{\beta})\}\Big] \\ &= \frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^{N} E_\xi\Big[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})\Big]; \end{aligned}$$

this, together with 3.14, implies

$$\begin{aligned} P_{\xi\pi}\Big(\sup_{\boldsymbol{\beta}\in\mathscr{O}} H_n(\boldsymbol{\beta}) > \varepsilon\Big) &\leq P_{\xi\pi}\left(\frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^{N} E_\xi\left[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})\right] > \varepsilon\right) \\ &\to 0. \end{aligned} \tag{3.15}$$

Similarly, it can be shown that

$$P_{\xi\pi}\Big(\inf_{\boldsymbol{\beta}\in\mathscr{O}} H_n(\boldsymbol{\beta}) < -\varepsilon\Big) \to 0. \tag{3.16}$$

Now, since Θ is compact there exist $m < \infty$ open balls $\mathscr{O}_j$ such that $\Theta \subset \bigcup_{j=1}^m \mathscr{O}_j$; then

$$\begin{aligned} P_{\xi\pi}\Big(\sup_{\boldsymbol{\beta}\in\Theta} |H_n(\boldsymbol{\beta})| > \varepsilon\Big) &\leq P_{\xi\pi}\Big(\sup_{\boldsymbol{\beta}\in\bigcup_{j=1}^m \mathscr{O}_j} |H_n(\boldsymbol{\beta})| > \varepsilon\Big) \\ &\leq \sum_{j=1}^{m} P_{\xi\pi}\Big(\sup_{\boldsymbol{\beta}\in\mathscr{O}_j} |H_n(\boldsymbol{\beta})| > \varepsilon\Big) \\ &\to 0, \end{aligned}$$

by 3.15 and 3.16.

Therefore, we get the desired result:

$$\sup_{\boldsymbol{\beta}\in\Theta} |H_n(\boldsymbol{\beta})| = \sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{N}\sum_s w_i\psi_i(Y_i,\boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^{N} E_\xi[\psi_i(Y_i,\boldsymbol{\beta})]\right| = \sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{N}s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right| \xrightarrow{p} 0.$$

□

Proof of Theorem 3.1. We carry out this proof in two cases:

CASE 1: Θ is a compact subset of $\mathbb{R}^p$
The following inequality holds:

$$\begin{aligned} \left|\frac{1}{N}s_n(\boldsymbol{\beta})\right| &= \left|\Delta_N(\boldsymbol{\beta}) + \frac{1}{N}s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right| \\ &\geq \left|\Delta_N(\boldsymbol{\beta})\right| - \left|\frac{1}{N}s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right|; \end{aligned}$$

therefore, making use of the lemma, for any $\varepsilon > 0$,

$$\begin{aligned}
\inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\varepsilon}\left|\frac{1}{N}s_n(\boldsymbol{\beta})\right| &\geq \inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\varepsilon}\left\{\left|\Delta_N(\boldsymbol{\beta})\right| - \left|\frac{1}{N}s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right|\right\} \\
&\geq \inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\varepsilon}\left|\Delta_N(\boldsymbol{\beta})\right| - \sup_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\varepsilon}\left|\frac{1}{N}s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right| \\
&\geq \inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\varepsilon}\left|\Delta_N(\boldsymbol{\beta})\right| - \sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{N}s_n(\boldsymbol{\beta}) - \Delta_N(\boldsymbol{\beta})\right| \\
&= \inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\varepsilon}\left|\Delta_N(\boldsymbol{\beta})\right| + o_p(1). \qquad (3.17)
\end{aligned}$$

By assumption 3, the right-hand side of 3.17 is strictly greater than zero in probability, as $n \to \infty$; and therefore, for any $\varepsilon > 0$, as $n \to \infty$,

$$P_{\xi\pi}\left(\inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\varepsilon}\left|\frac{1}{N}s_n(\boldsymbol{\beta})\right| > 0\right) \to 1.$$

Since, by assumption 4, $\hat{\boldsymbol{\beta}}_n$ solves $s_n(\boldsymbol{\beta}) = 0$ for $\boldsymbol{\beta} \in \Theta$, this limit implies that, for any $\varepsilon > 0$, $|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0| \leq \varepsilon$ in probability. This means that $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$.

CASE 2: Θ is any subset of $\mathbb{R}^p$

By assumption 5. in the theorem, for any $\varepsilon > 0$, there is an $M > 0$ such that $P_{\xi\pi}(||\hat{\boldsymbol{\beta}}_n|| \leq M) > 1-\varepsilon$ for all n. The result follows from case 1 by considering the closure of $\Theta \cap \{\boldsymbol{\beta} : ||\boldsymbol{\beta}|| \leq M\}$ as the parameter space. Let Θ^* be the clousure of $\Theta \cap \{\boldsymbol{\beta} : ||\boldsymbol{\beta}|| \leq M\}$. Then, for any $\delta > 0$,

$$\begin{aligned}
P_{\xi\pi}(||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| > \delta) &= P_{\xi\pi}(||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| > \delta,\ ||\hat{\boldsymbol{\beta}}_n|| \leq M) + P_{\xi\pi}(||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| > \delta,\ ||\hat{\boldsymbol{\beta}}_n|| > M) \\
&\leq P_{\xi\pi}(||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| > \delta,\ ||\hat{\boldsymbol{\beta}}_n|| \leq M) + P_{\xi\pi}(||\hat{\boldsymbol{\beta}}_n|| > M) \\
&< P_{\xi\pi}(||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| > \delta,\ ||\hat{\boldsymbol{\beta}}_n|| \leq M) + \varepsilon \\
&\leq P_{\xi\pi}(||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| > \delta,\ \hat{\boldsymbol{\beta}}_n \in \Theta^*) + \varepsilon \\
&= P_{\xi\pi}(\hat{\boldsymbol{\beta}}_n \in \Theta^*) P_{\xi\pi}(||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| > \delta \mid \hat{\boldsymbol{\beta}}_n \in \Theta^*) + \varepsilon \\
&\leq P_{\xi\pi}(||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| > \delta \mid \hat{\boldsymbol{\beta}}_n \in \Theta^*) + \varepsilon \\
&\leq 2\varepsilon,
\end{aligned}$$

where the last line is due to the fact that Θ^* is compact and then Case 1 applies. So that $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$. □

Chapter 4

Analysis of Longitudinal Surveys with Missing Responses

4.1 The Weighted Pseudo-GEE under a Model for Response Probabilities

The methods in section 2.2.3 apply to general studies with wave nonresponse, either monotone or intermittent. It turns out that if there are no time-dependent covariates, the two situations: wave nonresponse and missing response, are equivalent, since all time invariant covariates have been observed at wave 1. So, WGEE is also applicable in the case of missing responses with no time-dependent covariates. Nonetheless, even if there are (observed) time-dependent covariates, the case of missing response can be treated as a wave nonresponse situation, and the WGEE method can be applied. In this case one would ignore, for the main model, the information provided by the covariates whenever the response is missing, but can still use it for the nonresponse models. We now briefly show how the WGEE method of section 2.2.3 applies to longitudinal *surveys* with missing responses.

We assume that the superpopulation satisfies model ξ of section 2.2.3, and the finite population is a random sample of N elements from model ξ. From this finite population we select a (complex) sample s of n subjects, by means of design π; and each subject i has associated with it the (cycle 1's longitudinal) survey weight w_i. Some subjects drop out of the study at a given cycle. We define the response indicator variable R_{it} as 1 if subject i is observed at time t, and 0 otherwise. We assume that the response mechanism is MAR as described in equation 2.6, and that p_{it} follows the logistic regression model 2.7, which is indexed by λ. Recall that p_{it} is the probability that unit i responds at time t, having responded at time $t-1$ (and given its observed past). Also, recall that we use an estimate of p_{it} to adjust the observations from the respondents at time t to account for the nonrespondents at that time.

The parameters λ of the response model can be estimated by maximising the partial pseudo-likelihood

$$l(\lambda) = \sum_s w_i l_i(\lambda) = \sum_{i \in s} w_i \sum_{t=1}^{T} \log \left\{ p_{it}(\lambda)^{R_{it}} [1 - p_{it}(\lambda)]^{1-R_{it}} \right\}^{R_{i(t-1)}}. \tag{4.1}$$

We call $\hat{\lambda}$ such estimator and let $\pi_{it}(\lambda)$, $\pi_{it}(\hat{\lambda})$, $\Delta_i(\lambda)$, and $\Delta_i(\hat{\lambda})$ defined as in the "Monotone Missingness" subsection of section 2.2.3.

In the weighted GEE approach of section 2.2.3, the weights refer to the inverse of the response probabilities. With a complex survey sample, we have another set of weights, namely w_i. These are either the basic design weights or calibrated ones. Our proposed weighted pseudo-GEE method takes into account both sets of weights. First, if we follow the approach of Shao and Steel (1999) by dividing the finite population into two strata, one for respondents and the other for nonrespondents, then the weighted GEE, at the finite population level, would be

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \Delta_i(\hat{\lambda}_N)(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{4.2}$$

where $\hat{\lambda}_N$ is the partial pseudo maximum likelihood estimator of λ obtained from the "census" version of 4.1, and $\Delta_i(\lambda) = \text{diag}[R_{it}/\pi_{it}(\lambda)]$. Note that the weight in matrix $\Delta_i(\lambda)$ for any unobserved component of $\boldsymbol{y}_i$ is zero. In equation 4.2, $\boldsymbol{\mu}_i$ is a function of $\boldsymbol{\beta}$ and we call the solution for it B.

We propose a sample-based weighted GEE, termed as the Weighted Pseudo-GEE (WPGEE), using a Horvitz-Thompson type of estimator to the left hand side of 4.2, given by

$$\sum_{s} w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \Delta_i(\hat{\lambda})(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{4.3}$$

This equation is approximately unbiased under the joint randomization imposed by the superpopulation model, sampling design, and (MAR) response mechanism, $\xi\pi R$. The WPGEE estimator, $\hat{\boldsymbol{\beta}}_n$, is the solution to 4.3.

The joint variance of $\hat{\boldsymbol{\beta}}_n$ is given by

$$\begin{aligned} \text{Var}_{\xi\pi R}(\hat{\boldsymbol{\beta}}_n) &= \text{Var}_{\xi R}[E_\pi(\hat{\boldsymbol{\beta}}_n)] + E_{\xi R}[\text{Var}_\pi(\hat{\boldsymbol{\beta}}_n)] \\ &\doteq \text{Var}_{\xi R}(B) + E_{\xi R}[\text{Var}_\pi(\hat{\boldsymbol{\beta}}_n)] \end{aligned} \tag{4.4}$$

where

$$\text{Var}_{\xi R}(B) = \frac{1}{N}\Gamma^{-1} C (\Gamma^{-1})', \tag{4.5}$$

and Γ and C have the same form as in the "monotone missingness" part of section 2.2.3. The first component in 4.4, $\text{Var}_{\xi R}(B)$, is the model variance component and represents the variance in a census fit of the model; it is due to the fact that the N finite population data points scatter according to model ξ. The second component of 4.4, $E_{\xi R}[\text{Var}_\pi(\hat{\boldsymbol{\beta}}_n)]$, the design variance component, represents the additional variance contributed by sampling of the finite population.

We can estimate the joint variance of $\hat{\boldsymbol{\beta}}_n$ in 4.4 by estimating the two components separately, as

$$\hat{\text{V}}\text{ar}_{\xi\pi R}(\hat{\boldsymbol{\beta}}_n) = \hat{\text{V}}\text{ar}_{\xi R}(B) + \hat{\text{V}}\text{ar}_\pi(\hat{\boldsymbol{\beta}}_n);$$

where $\hat{\mathrm{V}}\mathrm{ar}_{\pi}(\hat{\boldsymbol{\beta}}_n)$ is an estimator of the design variance of $\hat{\boldsymbol{\beta}}_n$, which can be calculated using the bootstrap technique; and

$$\hat{\mathrm{V}}\mathrm{ar}_{\xi R}(B) = \frac{1}{\hat{N}}\hat{\Gamma}^{-1}\tilde{C}(\hat{\Gamma}^{-1})';$$

where $\hat{N} = \sum_s w_i$,

$$\hat{\Gamma} = \frac{1}{\hat{N}}\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n}\hat{V}_i^{-1}\Delta_i(\hat{\lambda})\frac{\partial \boldsymbol{\mu}_i}{\partial \hat{\boldsymbol{\beta}}_n},$$

$$\tilde{C} = \hat{I} - \hat{J}\hat{\Omega}\hat{J}',$$

$$\hat{I} = \frac{1}{\hat{N}}\sum_s w_i \Big[\frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n}\hat{V}_i^{-1}\Delta_i(\hat{\lambda})(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)\Big]^{\otimes 2},$$

$$\hat{J} = \frac{1}{\hat{N}}\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n}\hat{V}_i^{-1}\Lambda_i(\hat{\lambda}, \hat{\boldsymbol{\beta}}_n)\frac{\partial \pi_i(\lambda)}{\partial \hat{\lambda}'},$$

$\Lambda_i(\hat{\lambda}, \hat{\boldsymbol{\beta}}_n)$ and $\pi_i(\hat{\lambda})$ have the same form, and $\hat{\Omega}$ is the inverse of the observed information from the partial pseudo likelihood 4.1.

Equation 4.3 can also be written as

$$\sum_{i\in s}\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}\begin{bmatrix} \frac{w_i R_{i1}}{\pi_{i1}(\hat{\lambda})} & & & \mathrm{O} \\ & \frac{w_i R_{i2}}{\pi_{i2}(\hat{\lambda})} & & \\ & & \ddots & \\ \mathrm{O} & & & \frac{w_i R_{iT}}{\pi_{iT}(\hat{\lambda})} \end{bmatrix}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{4.6}$$

Equation 4.6 is "appealing" because it indicates that, for monotone longitudinal survey data, what WPGEE does is to adjust the original survey weight (for example cycle 1's longitudinal weight) for the inverse of the estimated probability that subject i responds at the given wave; and then weights the corresponding residual in $(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$ by this wave-specific "adjustment factor." Since Statistics Canada provides longitudinal weights for each cycle (as we show in Table 1.3), which are the cycle 1's longitudinal weights adjusted for nonresponse at the given cycle, it is likely that using Statistics Canada's cycle-specific weights in equation 4.6 produces similar results to the ones obtained by WGEE along with the estimation of the π_{it}'s (for example by logistic regression). Nonetheless, this approach is not directly applicable to the NLSCY because this survey has non-monotone (i.e. intermittent) missing patterns, and equation 4.6 is only applicable to monotone patterns.

With respect to intermittent patterns we proceed as follows. We make the same assumptions about the model ξ on the superpopulation, and sample design π; here some subjects who fail to respond at some cycle do come back to the survey at a later cycle. We define the response indicator variables R_{it}^* and R_{it}, and the vector R_i^* as in section 2.2.3.

We still assume that the response process satisfies equation 2.11, and that the response probabilities p_{it} follow the logistic regression model 2.12. Now we let $\hat{\lambda}$ solve the *weighted* estimating equation

$$\sum_s w_i S_{\lambda,i}(\lambda) = 0,$$

where $S_{\lambda,i}(\lambda)$ has the same form as in the "intermittent missingness" part of section 2.2.3. We let $\pi_i(r,\lambda)$, Φ_i, $A_i(\Phi,\hat{\lambda})$, and $\Delta_i(\hat{\lambda})$ as in section 2.2.3; but now we solve the following set of *weighted* equations to get the estimate, $\tilde{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$:

$$\sum_s w_i \left(\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \Delta_i(\hat{\lambda})(\boldsymbol{y}_i - \boldsymbol{\mu}_i) - \hat{\theta} A_i(\Phi,\hat{\lambda}) \right) = \mathbf{0}, \tag{4.7}$$

where $\hat{\theta} = \hat{\theta}_1 \hat{\theta}_2^{-1}$, and, if $U_i(\boldsymbol{\beta},\hat{\lambda}) = {}^{\partial \boldsymbol{\mu}_i'}/_{\partial \boldsymbol{\beta}} V_i^{-1} \Delta_i(\hat{\lambda})(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$,

$$\hat{\theta}_1 = \frac{1}{\hat{N}} \sum_s w_i e\big(U_i(\boldsymbol{\beta},\hat{\lambda}),\, S_{\lambda,i}(\hat{\lambda})\big) e\big(A_i(\Phi,\hat{\lambda}),\, S_{\lambda,i}(\hat{\lambda})\big)',$$

$$\hat{\theta}_2 = \frac{1}{\hat{N}} \sum_s w_i e\big(A_i(\Phi,\hat{\lambda}),\, S_{\lambda,i}(\hat{\lambda})\big) e\big(A_i(\Phi,\hat{\lambda}),\, S_{\lambda,i}(\hat{\lambda})\big)',$$

and $e(P_i,Q_i)$ is the residual for subject i from the *weighted* multivariate regression of the vectors P_i on the vectors Q_i.

Under 2.11, and provided the model for p_{it} is correctly specified, equation 4.7 has a root $\tilde{\boldsymbol{\beta}}$ that is consistent for $\boldsymbol{\beta}$ jointly under model, design, and response mechanism. This can be seen from the following argument;

$$\begin{aligned} E_{\xi\pi R}(\tilde{\boldsymbol{\beta}}) &= E_{\xi R}[E_\pi(\tilde{\boldsymbol{\beta}})] \\ &\doteq E_{\xi R}(B) \qquad (4.8) \\ &\doteq \boldsymbol{\beta}. \qquad (4.9) \end{aligned}$$

The estimating equations involved in 4.7 are approximately unbiased for those involved in the corresponding "census" equation, which typically implies that $\tilde{\boldsymbol{\beta}}$ is also approximately unbiased for B, which leads to 4.8. For a proof of 4.9 see Robins et al. (1995). Additionally, $\tilde{\boldsymbol{\beta}}$ is unique with probability approaching to one and asymptotically normal, under mild regularity conditions.

The joint variance of $\tilde{\boldsymbol{\beta}}$ is given by

$$\begin{aligned} \mathrm{Var}_{\xi\pi R}(\tilde{\boldsymbol{\beta}}) &= \mathrm{Var}_{\xi R}[E_\pi(\tilde{\boldsymbol{\beta}})] + E_{\xi R}[\mathrm{Var}_\pi(\tilde{\boldsymbol{\beta}})] \\ &\doteq \mathrm{Var}_{\xi R}(B) + E_{\xi R}[\mathrm{Var}_\pi(\tilde{\boldsymbol{\beta}})], \qquad (4.10) \end{aligned}$$

where

$$\mathrm{Var}_{\xi R}(B) = \frac{1}{N} \Gamma^{-1} \mathrm{Var}\big[e\big(U_i(\boldsymbol{\beta},\hat{\lambda}), B_i\big)\big] (\Gamma^{-1})', \tag{4.11}$$

and Γ and B_i' have the same form as in the intermittent missingness part of section 2.2.3. The first component in 4.10, $\mathrm{Var}_{\xi_R}(B)$, is the model variance component; and the second component, $E_{\xi_R}[\mathrm{Var}_\pi(\tilde{\boldsymbol{\beta}})]$, is the design variance component.

We can estimate the joint variance of $\tilde{\boldsymbol{\beta}}$ in 4.10 by estimating the two components separately, as

$$\hat{\mathrm{V}}\mathrm{ar}_{\xi\pi R}[\tilde{\boldsymbol{\beta}}] = \hat{\mathrm{V}}\mathrm{ar}_{\xi_R}(B) + \hat{\mathrm{V}}\mathrm{ar}_\pi(\tilde{\boldsymbol{\beta}});$$

where $\hat{\mathrm{V}}\mathrm{ar}_\pi(\tilde{\boldsymbol{\beta}})$ is an estimator of the design variance of $\tilde{\boldsymbol{\beta}}$, which can be estimated using usual design-based variance estimation techniques; and

$$\hat{\mathrm{V}}\mathrm{ar}_{\xi_R}(B) = \frac{1}{\hat{N}^2}\hat{\Gamma}^{-1}\sum_s w_i e\big(U_i(\tilde{\boldsymbol{\beta}},\hat{\lambda}),\hat{B}_i\big)(\hat{\Gamma}^{-1})';$$

where $\hat{B}_i'$ takes the same form as in the intermittent missingness part of section 2.2.3, $\hat{\Gamma}$ and $\tilde{C}$ have the same form as before, and

$$\hat{I} = \frac{1}{\hat{N}}\sum_s w_i\Big[\frac{\partial\boldsymbol{\mu}_i'}{\partial\tilde{\boldsymbol{\beta}}}\hat{V}_i^{-1}\Delta_i(\hat{\lambda})(\boldsymbol{y}_i-\hat{\boldsymbol{\mu}}_i)\Big]^{\otimes 2},$$

$$\hat{J} = \frac{1}{\hat{N}}\sum_s w_i\frac{\partial\boldsymbol{\mu}_i'}{\partial\tilde{\boldsymbol{\beta}}}\hat{V}_i^{-1}\Lambda_i(\hat{\lambda},\tilde{\boldsymbol{\beta}})\frac{\partial\pi_i(\lambda)}{\partial\hat{\lambda}'},$$

$\Lambda_i(\hat{\lambda},\tilde{\boldsymbol{\beta}})$ and $\pi_i(\hat{\lambda})$ have the same form, and $\hat{\Omega}$ is the inverse of the observed information from the partial pseudo likelihood 4.1.

Equations 4.4 and 4.10 are somewhat shaky even though the estimators $\hat{\boldsymbol{\beta}}_n$ and $\tilde{\boldsymbol{\beta}}$ are consistent for B under the design, i.e. $E_\pi(\hat{\boldsymbol{\beta}}_n) \doteq B$ and $E_\pi(\tilde{\boldsymbol{\beta}}) \doteq B$. The real issue here is that both $\hat{\boldsymbol{\beta}}_n$ and $\tilde{\boldsymbol{\beta}}$ have a bias which may not be of order $o(1/\sqrt{n})$, so bias becomes not negligible in deriving $\mathrm{Var}_{\xi_R}[E_\pi(\hat{\boldsymbol{\beta}}_n)]$ or $\mathrm{Var}_{\xi_R}[E_\pi(\tilde{\boldsymbol{\beta}})]$. However, if the sampling fraction is small, and $\mathrm{Var}_{\xi_R}[E_\pi(\hat{\boldsymbol{\beta}}_n)]$ and $\mathrm{Var}_{\xi_R}[E_\pi(\tilde{\boldsymbol{\beta}})]$ have the usual order of $1/N$, then, these model variance components (in 4.4 and 4.10) are negligible and can be dropped; as at the end of section 3.2.

4.2 Pseudo-GEE Method under Hot-deck Imputation

Large scale survey datasets are analyzed by different researchers, with different objectives of analyses. It is common practice nowadays for survey organizations, such as Statistics Canada, to produce a "complete" dataset with missing values imputed by appropriate methods. One advantage of using a common imputed data file is that different analyses can be compared with each other, and some internal consistency can be preserved.

For the rest of the thesis we assume that all the covariates are observed for all individuals at all times. Whereas the variable of interest, or response variable, is missing for some subjects and times. For some subjects we may have observed the response variable at all times. These subjects do not require any imputation and their actual observations are used in all the estimation procedures. Other subjects may have the response variable observed at some times

but not others. For these subjects we impute the response variable only at those times when it is missing; but use their actual observations when it is not. And for other subjects we may need to impute the response variable at all times.

We assume that our main model contains only categorical or ordinal covariates. If there *are* continuous covariates, for the time being, we ignore those in the imputation process. One may also categorize any continuous covariates thought to influence the missing mechanism if there is little or no loss in doing so.

At each cycle we divide all the subjects according to the cross-classified cells by covariates. Each cell in this classification is an imputation class. Within each cycle and imputation class we select a "donor," from among the respondents, for each nonrespondent, and replace the missing value of the response variable by the one from the donor. After we have filled-in all the missing responses at all times, we obtain a completed dataset.

With the completed dataset we are able to apply the usual GEE methodology, and standard software, to get the necessary point estimates. However, if we leave the correlation structure unspecified, most software packages will use method of moments to estimate the correlation parameters, using the completed (imputed) dataset. There might be a loss of efficiency by doing so; this requires further investigation. To avoid any possible loss of efficiency, in this thesis we leave the correlation structure unspecified *but* to estimate a given correlation parameter, say between times t and t', we use only those subjects who have *observed* responses at *both* times t and t' (see chapter 5 for more detail, in particular equations 5.3 and 5.8).

Nonetheless, this does not mean, automatically, that these estimators have good properties; or that standard measures of variability can be used. Therefore we need to address their properties.

Here we make the same assumptions about the model, ξ, and the sample, s, as in section 3.2. As was mentioned above, we only consider cases where all the covariates are either categorical or ordinal. We also assume that covariates which affect the response-nonresponse mechanism are all included in the superpopulation model ξ, so we can form imputation classes using covariates from the model ξ only. The total number of imputation classes is determined by the number of cross-classified cells using all covariates. Under such a formulation of imputation classes, the MAR assumption is satisfied.

By doing this, the model expectation of all responses within the same imputation class will be constant, whether the response is observed or missing. This condition may not be imperative in practice but is the assumption we need to facilitate the proof of the consistency of the resulting estimators of the model parameters.

The classification into imputation classes of all the subjects at each time, according to the cross-classification by covariates, leads to r_{jg} respondents and $m_{jg} = n_{jg} - r_{jg}$ nonrespondents, at time j, in cell g of the cross-classification; where n_{jg} is the number of selected individuals falling into cell g, $g = 1,2,\cdots,G$ at time j; and $G = (c_1 \times c_2 \times \cdots \times c_p)$, and c_k is the number of categories of the k-th covariate.

It may occur that, for some cycle(s), some cells contain no sampled elements at all; the estimation method still works out in this case. Nevertheless, a problem does arise when some cell has one or more missing values yet not any respondents; in such an instance there would be no available donor for some nonrespondent(s). In the simulations studies of chapter 5 we

discard any sample in which this situation occurs. In practice, one would have to deal with it in some other way; for example one could collapse some neighbouring cells or merge some categories of some covariate(s).

If we define imputation classes based on covariates used in the model, and if some of these covariates are time-varying, we could end up with different imputation classes at different time points (i.e. waves). This may not impose a problem for practical applications, but it does cause notational problems for our theoretical developments here. In what follows, we assume that our imputation classes are formulated in a refined way, such that any donor set for a particular missing response will not cut across imputation classes at different time points.

We then have that the finite population U can be partitioned, into "imputation classes", as $U = U_1 \cup U_2 \cup \cdots U_G = \bigcup_{g=1}^{G} U_g$; so that we can write the "census" estimating function as

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \sum_{g=1}^{G} \sum_{i \in U_g} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i);$$

and at the sample level, $s = \cup_{g=1}^{G} s_g$, and $\sum_s {}^{\partial \boldsymbol{\mu}_i}/_{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \sum_{g=1}^{G} \sum_{i \in s_g} {}^{\partial \boldsymbol{\mu}_i}/_{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)$. Now, following the framework used by Shao and Steel (1999), we consider that the finite population can also be partitioned into "respondents" and "nonrespondents," as $U = U_r \cup U_m$, where $U_r = \cup_{g=1}^{G} U_{gr}$, and U_{gr} is the set of "responding" units in cell g of the finite population. Similarly for U_m.

The double summation should carry over for all technical arguments. It is easy to see, as demonstrated in the proof of property 4.1 below, that all key arguments come from within each imputation class, and then sum up over all classes. So, without loss of generality, we will proceed as if we only have one imputation class, for simplicity of notation.

We consider two commonly used hot-deck imputation procedures; the unweighted and the weighted hot-deck imputation. Under "unweighted hot-deck," for each time j and each cell g we draw a simple random sample with replacement of size m_{jg}, from among the r_{jg} respondents, which serve as donors for the m_{jg} nonrespondents. The selection of the m_{jg} donors is done with equal probabilities among the r_{jg} respondents. For "weighted hot-deck," on the other hand, the selection of donors is carried out with probabilities proportional to the respondent's weight, w_i.

We assume that the model (ξ), sampling (π), response (R), and imputation (I) mechanisms are unconfounded. According to Brick et al. (2004), this means that after conditioning on all auxiliary variables, the distribution of the variable Y is independent of the other three mechanisms. This allows the interchange of the model expectation and the other three.

Let $\boldsymbol{y}_i^*$ be the vector of all the observed responses, $\boldsymbol{y}_i^{\text{obs}} = \boldsymbol{y}_i^{\text{o}}$, and imputed responses, $\boldsymbol{y}_i^{\text{imp}} = \boldsymbol{y}_i^{\text{I}}$, for subject i, organized as $\boldsymbol{y}_i^* = \left((\boldsymbol{y}_i^{\text{obs}})', (\boldsymbol{y}_i^{\text{imp}})'\right)' = \left((\boldsymbol{y}_i^{\text{o}})', (\boldsymbol{y}_i^{\text{I}})'\right)'$. We also sort the components of $\boldsymbol{\mu}_i$ correspondingly, as $\boldsymbol{\mu}_i = \left((\boldsymbol{\mu}_i^{\text{obs}})', (\boldsymbol{\mu}_i^{\text{mis}})'\right)' = \left((\boldsymbol{\mu}_i^{\text{o}})', (\boldsymbol{\mu}_i^{\text{M}})'\right)'$.

We estimate $\boldsymbol{\beta}$ applying the GEE method to the "filled-in" or "completed" dataset; i.e. we use $\boldsymbol{y}_i^*$ as the response variable for subject i. We solve the following set of equations

$$U_n^*(\boldsymbol{\beta}) = \sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{4.12}$$

where we sort the T_i columns of $\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta}$, the T_i rows and columns of V_i, and the T_i elements of $\boldsymbol{\mu}_i$ according to the corresponding ordering of the vector $\boldsymbol{y}_i^*$. V_i is a working variance-covariance matrix of Y_i, composed as $V_i = A_i^{1/2}\mathbf{R}_i A_i^{1/2}$, where A_i is a $T_i \times T_i$ diagonal matrix with $\mathrm{Var}_\xi[Y_{ij}] = \phi\, v(\mu_{ij})$ as the j-th diagonal element, and $\mathbf{R}_i$ is a working correlation matrix for Y_i.

We now examine asymptotic properties of the pseudo-GEE estimator $\hat{\boldsymbol{\beta}}$, which is the solution to 4.12, under imputation for missing response variable, using either weighted or unweighted hot-deck imputation method [1]. The consistency of $\hat{\boldsymbol{\beta}}$ under the joint $\xi\pi RI$ randomization is established in Theorem 4.2, below. Proofs are given in section 4.5.

Property 4.1. *The estimating function $U_n^*(\boldsymbol{\beta})$, on the left hand side of equation 4.12 is unbiased with respect to the model, design, response, and imputation mechanisms if these are unconfounded.*

Proof of Property 4.1.

$$\begin{aligned}
E_{\xi\pi RI}[U_n^*(\boldsymbol{\beta})] &= E_{\xi\pi RI}\left[\sum_s w_i \frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}} V_i^{-1}(Y_i^* - \boldsymbol{\mu}_i)\right] \\
&= E_{\xi\pi RI}\left[\sum_{g=1}^{G}\sum_{s_g} w_i \frac{\partial\boldsymbol{\mu}_{(g)}'}{\partial\boldsymbol{\beta}} V_i^{-1}(Y_i^* - \boldsymbol{\mu}_{(g)})\right] \\
&= E_{\pi RI}\left[\sum_{g=1}^{G}\sum_{s_g} w_i \frac{\partial\boldsymbol{\mu}_{(g)}'}{\partial\boldsymbol{\beta}} V_i^{-1}(E_\xi[Y_i^*] - \boldsymbol{\mu}_{(g)})\right] \\
&= E_{\pi RI}\left[\sum_{g=1}^{G}\sum_{s_g} w_i \frac{\partial\boldsymbol{\mu}_{(g)}'}{\partial\boldsymbol{\beta}} V_i^{-1}(\boldsymbol{\mu}_{(g)} - \boldsymbol{\mu}_{(g)})\right] \\
&= \mathbf{0},
\end{aligned}$$

where $\boldsymbol{\mu}_{(g)} = g^{-1}(\mathbf{X}_{(g)}'\boldsymbol{\beta})$ is the model-mean response in class g, and $\mathbf{X}_{(g)}'$ is a typical covariate vector value in that class. Also, in class g, $E_\xi[Y_i^*] = \boldsymbol{\mu}_{(g)}$ because inside that class, all elements (either observed or missing) have the same ξ-expectation. □

NOTE that the previous result remains valid for any combination of model and imputation strategy satisfying $E_{\xi I}[Y_i^*] = \boldsymbol{\mu}_i$. Besides the hot-deck imputation method described above, other examples of this are mean imputation within cells (for only-categorical covariates), and regression imputation (when the model ξ is linear). The latter is also applicable with continuous covariates.

In what follows we assume that the selection of a certain respondent i, as donor, in the hot-deck imputation procedure, is carried out with probability proportional to a "weight" τ_i. This τ_i corresponds exactly to w_i if weighted hot-deck is used, and to 1 if we use unweighted

[1] For the remainder of this thesis we call $\hat{\boldsymbol{\beta}}$ the *actual estimator used* (obtained with the imputed dataset), i.e. the solution to $U_n^*(\boldsymbol{\beta}) = \mathbf{0}$; $\hat{\boldsymbol{\beta}}_n$ the *conceptual* estimator obtained in the ideal situation of 100% response rate; B the *conceptual* estimator obtained in a census situation; and $\boldsymbol{\beta}$ the (superpopulation) *parameter* of interest.

hot-deck. We also assume $r/n = O(1)$; so that, for instance, $O(1/\sqrt{r}) = O(1/\sqrt{n})$. The case with $r/n = 1 + o(1)$ is trivial, since it implies negligible missing fraction for large samples.

Theorem 4.2. *Let $s_n(\boldsymbol{\beta}) = \sum_s w_i \psi_i(Y_i, \boldsymbol{\beta})$, $s_n^*(\boldsymbol{\beta}) = \sum_s w_i \psi_i(Y_i^*, \boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \Theta \subset \mathbb{R}^p$, $\psi_i(Y_i, \boldsymbol{\beta})$ is a function from $\mathbb{R}^{T_i} \times \Theta$ to $\mathbb{R}^p$, and Y_i^* is the vector Y_i with missing values imputed by the hot-deck method (either weighted or unweighted); let $\boldsymbol{\beta}_0 \in \Theta$ be such that $E_{\xi\pi}[s_n(\boldsymbol{\beta}_0)] = 0$; let $h_i(Y_i) = \sup_{\boldsymbol{\beta}\in\Theta} \|\psi_i(Y_i, \boldsymbol{\beta})\|$, $i = 1, 2, \ldots$, where $\|\cdot\|$ is the usual $\mathscr{L}_1$ norm; and let $\Delta_N^*(\boldsymbol{\beta}) = E_{\xi\pi RI}[N^{-1} s_n^*(\boldsymbol{\beta})]$. Suppose that*

1. *$\sup_i E_\xi |h_i(Y_i)|^2 < \infty$ and $\sup_i E_\xi ||Y_i|| < \infty$;*
2. *For any $c > 0$ and sequence $\{y_i\}$ satisfying $\|y_i\| \leq c$, the sequence of functions $\{g_i(\boldsymbol{\beta}) = \psi_i(y_i, \boldsymbol{\beta})\}$ is equicontinuous on any open subset of Θ;*
3. *The function $\Delta_N(\boldsymbol{\beta}) = E_{\xi\pi}[N^{-1} s_n(\boldsymbol{\beta})]$ has the property that $\inf_{|\boldsymbol{\beta}-\boldsymbol{\beta}_0|>\varepsilon} |\Delta_N(\boldsymbol{\beta})| > 0$ for any $\varepsilon > 0$;*
4. *There exists a $\hat{\boldsymbol{\beta}} \in \Theta$ that is solution to $s_n^*(\boldsymbol{\beta}) = 0$, i.e. $\hat{\boldsymbol{\beta}}$ is the pseudo-GEE estimator of $\boldsymbol{\beta}$ such that $s_n^*(\hat{\boldsymbol{\beta}}) = 0$;*
5. *$\hat{\boldsymbol{\beta}} = O_p(1)$;*
6. *The design weights w_i satisfy $N^{-1}\sum_s w_i Z_i - N^{-1}\sum_{i=1}^N Z_i = O_p(1/\sqrt{n})$ for any variable Z such that $N^{-1}\sum_{i=1}^N Z_i^2 = O(1)$;*

then $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, where "p" denotes in probability with respect to the model ξ, the sampling design π, the response mechanism R, and the imputation mechanism I.

The proof of Theorem 4.2 is given in section 4.5.1.

4.3 Variance Estimation under Hot-deck Imputation

In this section we develop linearization variance estimators for the pseudo-GEE estimator $\hat{\boldsymbol{\beta}}$ under imputation for missing responses. The general strategy is to break down the total variance into various variance components, and then estimate them piece by piece.

There are four randomization processes and each of them contributes some amount of error. The error inherited by the model; this is produced because the N elements in the population are generated by model ξ. Then the error generated by the sampling mechanism π; which is basically due to the fact that we do not observe the whole finite population, but instead select only a portion of it to be observed. Additionally, not all elements in the sample are observed; only a part of it, which is generated by the response mechanism R, which we assume to be MAR. And finally the imputation error; this has two features. The imputation process I is usually a random mechanism, and if we repeated the imputation step, we would obtain a different set of imputed values. Additionally, even with deterministic imputation there is error in the imputed

values. Särndal (1992) points out that "the variance of an estimated total is increased by imputation, because imputation does not (except in truly exceptional circumstances) reproduce the true value y_k."

Särndal's observation may be interpreted in two ways. On the one hand, it should be clear that under *random* imputation, the variance of an estimated total using the imputed values should be higher than that of the complete case estimator. But even for deterministic imputation that seems to be the case. For example, in all the cases considered by Chen and Shao (2000) the variance of an estimated mean using data completed by nearest neighbour imputation is higher than the variance of the complete case estimator.

The other way of interpreting Särndal's remark is to compare the variance of the estimated total using imputed values to the variance of the conceptual estimator that one would obtain if there were no missing values. The variance of the former should be higher than that of the latter because of two reasons. Not only do the imputed values vary from sample to sample (whereas if there were 100% response rate they would not) but also they are sample-based (more specifically, respondent set-based).

We now introduce some notation. We let $\hat{\boldsymbol{\beta}}$ be the solution to equations 4.12. We call $\hat{\boldsymbol{\beta}}_n$ the vector that solves the following set of equations

$$U_n(\boldsymbol{\beta}) = \frac{1}{N}\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{4.13}$$

In other words, $\hat{\boldsymbol{\beta}}_n$ is the estimator vector one would obtain if all the elements selected in the sample s had been fully observed. And finally we let B be the solution to the equations

$$U_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}. \tag{4.14}$$

That is, B is the estimator of $\boldsymbol{\beta}$ obtained when the GEE method is applied to the whole finite population.

Now we can decompose the total error in the estimator $\hat{\boldsymbol{\beta}}$ as follows

$$\begin{aligned}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n) + (\hat{\boldsymbol{\beta}}_n - B) + (B - \boldsymbol{\beta}) \\ &= (\text{Imputation error}) + (\text{Sampling error}) + (\text{Model error}).\end{aligned}$$

We consider the practical situation for most complex longitudinal surveys where the sampling fraction is small or negligible, i.e. $n/N = o(1)$. We also assume that the usual $\sqrt{n}$ order applies to various estimators used in the decomposition of the total variance, i.e. $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n = O_p(1/\sqrt{r})$, $\hat{\boldsymbol{\beta}}_n - B = O_p(1/\sqrt{n})$, $B - \boldsymbol{\beta} = O_p(1/\sqrt{N}) = o_p(1/\sqrt{r})$; where r is the number of respondents, n is the sample size, and N is the finite population size. We can therefore ignore all terms involving $B - \boldsymbol{\beta}$. This leads to the following decomposition of the total variance:

$$\begin{aligned}V_{\text{Tot}} =& E_{\xi\pi RI}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \\ =& E_{\xi\pi RI}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)' + E_{\xi\pi RI}(\hat{\boldsymbol{\beta}}_n - B)(\hat{\boldsymbol{\beta}}_n - B)' + \left\{E_{\xi\pi RI}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n - B)'\right.\end{aligned}$$

$$
\begin{aligned}
&+E_{\xi\pi RI}(\hat{\boldsymbol{\beta}}_n - B)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)'\} + o(1/r) \\
=&V_{\text{Imp}} + V_{\text{Sam}} + \{C_{\text{Imp-Sam}} + C'_{\text{Imp-Sam}}\} + o(1/r),
\end{aligned} \tag{4.15}
$$

where $V_{\text{Imp}} = E_{\xi\pi R}V_{I(I)}$, $V_{I(I)} = E_I(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)'$; $V_{\text{Sam}} = E_\xi V_\pi$, $V_\pi = \text{Var}_\pi(\hat{\boldsymbol{\beta}}_n - B) = E_\pi(\hat{\boldsymbol{\beta}}_n - B)(\hat{\boldsymbol{\beta}}_n - B)'$; and $C_{\text{Imp-Sam}} = E_{\pi RI}C_{\xi(I\pi)}$, $C_{\xi(I\pi)} = E_\xi(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n - B)'$.

Asymptotic expansions for each of the terms involved in 4.15, which are used to construct a variance estimator, are given in Theorem 4.3, below; the proof will be given in section 4.5.2. Let

$$
H(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}, \text{ and } \hat{H}(\boldsymbol{\beta}) = \sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}},
$$

$\pi_i = 1/w_i$, $\Delta_{ii} = \pi_i(1-\pi_i)$, π_{ij} be the joint probability of inclusion of elements i and j, and $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$, for $i \neq j$.

Theorem 4.3. *Under conditions that lead to 4.15, the four leading variance-covariance components can be approximated as follows:*

(1) The imputation variance component:

$$
\begin{aligned}
V_{Imp} =&E_{\xi\pi R}V_{I(I)} = E_{\xi\pi R}E_I(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)' \\
\doteq&E_{\xi\pi R}\Big\{\big[\hat{H}(\hat{\boldsymbol{\beta}}_n)\big]^{-1}\Big[\sum_s w_i^2 \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & diag(s^2_{\tau r}) \end{pmatrix} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \hat{\boldsymbol{\beta}}_n} + \Big(\sum_{i\in s} w_i E_I(\mathbf{z}_i^*)\Big)^{\otimes 2}\Big]\big[\hat{H}(\hat{\boldsymbol{\beta}}_n)\big]^{-1}\Big\},
\end{aligned} \tag{4.16}
$$

where $s^2_{\tau r} = \sum_r \tau_j y_j^2/\sum_r \tau_j - (\sum_r \tau_j y_j/\sum_r \tau_j)^2$ *is the* τ*-weighted sample variance and* $\bar{y}_{\tau r} = \sum_r \tau_k y_j/\sum_r \tau_j$ *the* τ*-weighted mean of respondents in the given cycle, and*

$$
E_I(\mathbf{z}_i^*) = \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1} \begin{pmatrix} \mathbf{y}_i^o - \boldsymbol{\mu}_i^o \\ \bar{y}_{\tau r} - \boldsymbol{\mu}_i^M \end{pmatrix};
$$

(2) The sampling variance component:

$$
\begin{aligned}
V_{Sam} =&E_\xi V_\pi = E_\xi Var_\pi(\hat{\boldsymbol{\beta}}_n - B) \\
\doteq&E_\xi\Big\{\big[H(B)\big]^{-1}\Big[\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\Delta_{ij}}{\pi_i\pi_j}\mathbf{z}_i\mathbf{z}_j'\Big]\big[H(B)\big]^{-1}\Big\},
\end{aligned} \tag{4.17}
$$

where $\mathbf{z}_i = {}^{\partial \boldsymbol{\mu}_i'}/_{\partial B} V_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)$*; and*

(3) The cross, imputation-sampling, component:

$$
\begin{aligned}
&C_{Imp\text{-}Sam} = E_{\pi RI}C_{\xi(I\pi)} = E_{\pi RI}E_\xi(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n - B)' \\
&\doteq E_{\pi RI}\Bigg\{\big[\hat{H}(\boldsymbol{\beta})\big]^{-1} \times \Bigg[\sum_{i,j\in s}\sum_{i\neq j} w_i w_j \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ E_\xi(\mathbf{e}_i^I \mathbf{e}_j^{o\prime}) & \mathbf{0} \end{pmatrix} V_j^{-1} \frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{\beta}}
\end{aligned}
$$

$$-\sum_s w_i^2 \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ E_\xi(\mathbf{e}_i^M \mathbf{e}_i^{o\prime}) & E_\xi(\mathbf{e}_i^M \mathbf{e}_i^{M\prime}) \end{pmatrix} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \Bigg] \times \left[\hat{H}(\boldsymbol{\beta})\right]^{-1} \Bigg\}, \tag{4.18}$$

where $\mathbf{e}_i^o = \mathbf{y}_i^o - \boldsymbol{\mu}_i^o$ *and* $\mathbf{e}_i^I = \mathbf{y}_i^I - \boldsymbol{\mu}_i^M$ *are the "observed" and "imputed" parts of the error* $\mathbf{e}_i^* = \mathbf{y}_i^* - \boldsymbol{\mu}_i$*, respectively; and* $\mathbf{e}_i^M = \mathbf{y}_i^M - \boldsymbol{\mu}_i^M$ *is the "missing" part of the error* $\mathbf{e}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$*. Note that* $\mathbf{e}_i^o = \mathbf{y}_i^o - \boldsymbol{\mu}_i^o$ *is also the "observed" part of the error* $\mathbf{e}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$*.*

To estimate the total variance of $\hat{\boldsymbol{\beta}}$, in Theorem 4.3, we follow the approach used in Särndal (1992) and in Brick et al. (2004). We get (approximately) unbiased estimators of each of the three components, V_{Imp}, V_{Sam}, and $C_{\text{Imp-Sam}}$, and add them up to get an approximately unbiased estimator of V_{Tot}.

4.3.1 Estimation of Imputation Variance Component

For the imputation variance component, V_{Imp}, we can use a simple "plug-in" estimator of $V_{I(I)}$ (and V_{Imp}):

$$\begin{aligned} \hat{V}_{\text{Imp}} &= \hat{V}_{I(I)} \\ &= \left[\hat{H}(\hat{\boldsymbol{\beta}})\right]^{-1} \left[\sum_s w_i^2 \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \text{diag}(s_{\tau r}^2) \end{pmatrix} \hat{V}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\boldsymbol{\beta}}} + \left(\sum_{i \in s} w_i \widehat{E_I(\mathbf{z}_i^*)}\right)^{\otimes 2}\right] \left[\hat{H}(\hat{\boldsymbol{\beta}})\right]^{-1}, \end{aligned} \tag{4.19}$$

where $\widehat{E_I(\mathbf{z}_i^*)} = {}^{\partial \hat{\boldsymbol{\mu}}_i'}\!/_{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \begin{pmatrix} \mathbf{y}_i^o - \hat{\boldsymbol{\mu}}_i^o \\ \bar{y}_{\tau r} - \hat{\boldsymbol{\mu}}_i^M \end{pmatrix}$.

4.3.2 Estimation of Sampling Variance Component

With regard to the estimation of the sampling component, V_{Sam}, the first and last terms (the "bread") on the RHS of equation 4.17 do not involve the design and we can just estimate it with $\hat{H}(\hat{\boldsymbol{\beta}})$. On the other hand, the estimation of the term in the middle (the "meat") is not straightforward. This term is:

$$\text{Var}_\pi[U_n(B)] = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \frac{\Delta_{ij}}{\pi_i \pi_j} \mathbf{z}_i \mathbf{z}_j'.$$

According to Särndal et al. (1992), this can be unbiasedly estimated by

$$\frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij} \pi_i \pi_j} \mathbf{z}_i \mathbf{z}_j'.$$

However, even if all the elements in the sample s were fully observed, this estimator requires the knowledge of the joint probabilities of inclusion for every pair of sampled elements. Since

this is often unfeasible, an alternative estimator, which assumes the PSUs are selected with replacement, is

$$\frac{1}{n(n-1)}\sum_{k\in s}\Big[nw_k\boldsymbol{z}_k-\sum_{i\in s}w_i\boldsymbol{z}_i\Big]^{\otimes 2}=\frac{1}{n-1}\Big[(n-1)\sum_{k\in s}w_k^2\boldsymbol{z}_k\boldsymbol{z}_k'-\sum_{k,i\in s}\sum_{i\neq k}w_kw_i\boldsymbol{z}_k\boldsymbol{z}_i'\Big] \tag{4.20}$$

$$=\frac{1}{n-1}\Big[n\sum_{k\in s}w_k^2\boldsymbol{z}_k\boldsymbol{z}_k'-\Big(\sum_{i\in s}w_i\boldsymbol{z}_i\Big)^{\otimes 2}\Big] \tag{4.21}$$

where $\boldsymbol{z}_i={}^{\partial\boldsymbol{\mu}_i'}/_{\partial B}V_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)$ and $A^{\otimes 2}=AA'$. This estimator has a positive bias when the PSUs are selected without replacement. However, if the first stage sampling fraction is small this bias is negligible. We are going to use this estimator. Nonetheless, we cannot apply it directly since we have some missing y_{ij}'s (and then some missing terms in some $\boldsymbol{z}_i$'s). A naïve approach replaces any missing y_{ij} by the corresponding imputed value y_{ij}^{I}. This method underestimates the true variability. If we let $\boldsymbol{z}_i^*={}^{\partial\boldsymbol{\mu}_i'}/_{\partial B}V_i^{-1}(\boldsymbol{y}_i^*-\boldsymbol{\mu}_i)$, our naïve estimator of the "meat" would be

$$\frac{1}{n-1}\Big[n\sum_{k\in s}w_k^2\boldsymbol{z}_k^*\boldsymbol{z}_k^{*\prime}-\Big(\sum_{i\in s}w_i\boldsymbol{z}_i^*\Big)^{\otimes 2}\Big],$$

which has the following (model) bias:

$$\sum_{k\in s}w_k^2\frac{\partial\boldsymbol{\mu}_k'}{\partial B}V_k^{-1}\Big[\begin{pmatrix}\mathbf{0}&\mathbf{0}\\ \mathbf{0}&E_\xi(\boldsymbol{e}_k^{\mathrm{I}}\boldsymbol{e}_k^{\mathrm{I}\prime})\end{pmatrix}-\begin{pmatrix}\mathbf{0}&E_\xi(\boldsymbol{e}_k^{\mathrm{o}}\boldsymbol{e}_k^{\mathrm{M}\prime})\\ E_\xi(\boldsymbol{e}_k^{\mathrm{M}}\boldsymbol{e}_k^{\mathrm{o}\prime})&E_\xi(\boldsymbol{e}_k^{\mathrm{M}}\boldsymbol{e}_k^{\mathrm{M}\prime})\end{pmatrix}\Big]V_k^{-1}\frac{\partial\boldsymbol{\mu}_k}{\partial B}$$

$$-\frac{1}{n-1}\sum_{k,i\in s}\sum_{i\neq k}w_kw_i\frac{\partial\boldsymbol{\mu}_k'}{\partial B}V_k^{-1}\begin{pmatrix}\mathbf{0}&E_\xi(\boldsymbol{e}_k^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{I}\prime})\\ E_\xi(\boldsymbol{e}_k^{\mathrm{I}}\boldsymbol{e}_i^{\mathrm{o}\prime})&E_\xi(\boldsymbol{e}_k^{\mathrm{I}}\boldsymbol{e}_i^{\mathrm{I}\prime})\end{pmatrix}V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial B}, \tag{4.22}$$

where $\boldsymbol{e}_k^{\mathrm{I}}=\boldsymbol{y}_k^{\mathrm{I}}-\boldsymbol{\mu}_k^{\mathrm{M}}$, $\boldsymbol{e}_k^{\mathrm{o}}=\boldsymbol{y}_k^{\mathrm{o}}-\boldsymbol{\mu}_k^{\mathrm{o}}$, $\boldsymbol{e}_k^{\mathrm{M}}=\boldsymbol{y}_k^{\mathrm{M}}-\boldsymbol{\mu}_k^{\mathrm{M}}$. The order of magnitude of the double summation term in 4.22 is unclear. But simulation studies reported in chapter 5 reveal that its contribution is very minor and may be omitted for simplicity; at least for simple random sampling and stratified random sampling.

In order to obtain an estimator of this bias we must find estimators of the quantities $E_\xi(\boldsymbol{e}_k^{\mathrm{I}}\boldsymbol{e}_k^{\mathrm{I}\prime})$, $E_\xi(\boldsymbol{e}_k^{\mathrm{o}}\boldsymbol{e}_k^{\mathrm{M}\prime})$, $E_\xi(\boldsymbol{e}_k^{\mathrm{M}}\boldsymbol{e}_k^{\mathrm{M}\prime})$, $E_\xi(\boldsymbol{e}_k^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{I}\prime})$, and $E_\xi(\boldsymbol{e}_k^{\mathrm{I}}\boldsymbol{e}_i^{\mathrm{I}\prime})$. Since $\boldsymbol{y}_k^{\mathrm{I}}$ and $\boldsymbol{y}_k^{\mathrm{o}}$ are available for all elements $k\in s$ (the former are imputed values and the latter are observed ones), we can estimate $E_\xi(\boldsymbol{e}_k^{\mathrm{I}}\boldsymbol{e}_k^{\mathrm{I}\prime})$, $E_\xi(\boldsymbol{e}_k^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{I}\prime})$, and $E_\xi(\boldsymbol{e}_k^{\mathrm{I}}\boldsymbol{e}_i^{\mathrm{I}\prime})$ with $\boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_k^{\mathrm{I}\prime}$, $\boldsymbol{r}_k^{\mathrm{o}}\boldsymbol{r}_i^{\mathrm{I}\prime}$, and $\boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_i^{\mathrm{I}\prime}$, respectively; where $\boldsymbol{r}_k^{\mathrm{o}}=\boldsymbol{y}_k^{\mathrm{o}}-\hat{\boldsymbol{\mu}}_k^{\mathrm{o}}$ and $\boldsymbol{r}_k^{\mathrm{I}}=\boldsymbol{y}_k^{\mathrm{I}}-\hat{\boldsymbol{\mu}}_k^{\mathrm{M}}$ are the "observed" and "imputed" parts of the residual $\boldsymbol{r}_k=\boldsymbol{y}_k^*-\hat{\boldsymbol{\mu}}_k$.

On the other hand, $\boldsymbol{y}_k^{\mathrm{M}}$ is never observed -it is by definition "the missing part of $\boldsymbol{y}_k$",- so that $E_\xi(\boldsymbol{e}_k^{\mathrm{o}}\boldsymbol{e}_k^{\mathrm{M}\prime})$ and $E_\xi(\boldsymbol{e}_k^{\mathrm{M}}\boldsymbol{e}_k^{\mathrm{M}\prime})$ cannot be estimated in the same form. We can, nonetheless, make use of the matrix $\hat{V}_i$ for this purpose. This matrix is an estimator of $E_\xi(\boldsymbol{e}_i\boldsymbol{e}_i{}')=E_\xi(\boldsymbol{y}_i-\boldsymbol{\mu}_i)(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'$ for every i. Then, for an element k we can take the sub-matrix of $\hat{V}_i$ corresponding to the observed rows and missing columns of element k as an estimator of $E_\xi(\boldsymbol{e}_k^{\mathrm{o}}\boldsymbol{e}_k^{\mathrm{M}\prime})$; and the sub-matrix of $\hat{V}_i$ corresponding to missing rows and missing columns of element k as an estimator of $E_\xi(\boldsymbol{e}_k^{\mathrm{M}}\boldsymbol{e}_k^{\mathrm{M}\prime})$. The naïve estimator of V_π is then:

$$\hat{V}_\pi^*=\big[\hat{H}(\hat{\boldsymbol{\beta}})\big]^{-1}\Big[\frac{1}{n-1}\Big\{n\sum_{k\in s}w_k^2\boldsymbol{z}_k^*\boldsymbol{z}_k^{*\prime}-\Big(\sum_{i\in s}w_i\boldsymbol{z}_i^*\Big)^{\otimes 2}\Big\}\Big]\big[\hat{H}(\hat{\boldsymbol{\beta}})\big]^{-1},$$

whose bias can be estimated by

$$\big[\hat{H}(\hat{\boldsymbol{\beta}})\big]^{-1}\times\Bigg\{\sum_{k\in s}w_k^2\frac{\partial\hat{\boldsymbol{\mu}}_k'}{\partial\hat{\boldsymbol{\beta}}}\hat{V}_k^{-1}\Big[\begin{pmatrix}\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_k^{\mathrm{I}\prime}\end{pmatrix}-\begin{pmatrix}\mathbf{0}&\hat{V}_k^{(\mathrm{oM})}\\ \hat{V}_k^{(\mathrm{oM})\prime}&\hat{V}_k^{(\mathrm{MM})}\end{pmatrix}\Big]\hat{V}_k^{-1}\frac{\partial\hat{\boldsymbol{\mu}}_k}{\partial\hat{\boldsymbol{\beta}}}$$
$$-\frac{1}{n-1}\sum_{k,i\in s}\sum_{i\neq k}w_kw_i\frac{\partial\hat{\boldsymbol{\mu}}_k'}{\partial\hat{\boldsymbol{\beta}}}\hat{V}_k^{-1}\begin{pmatrix}\mathbf{0}&\boldsymbol{r}_k^{\mathrm{o}}\boldsymbol{r}_i^{\mathrm{I}\prime}\\ \boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_i^{\mathrm{o}\prime}&\boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_i^{\mathrm{I}\prime}\end{pmatrix}\hat{V}_i^{-1}\frac{\partial\hat{\boldsymbol{\mu}}_i}{\partial\hat{\boldsymbol{\beta}}}\Bigg\}\times\big[\hat{H}(\hat{\boldsymbol{\beta}})\big]^{-1},\quad(4.23)$$

where $\boldsymbol{r}_k^{\mathrm{o}}=\boldsymbol{y}_k^{\mathrm{o}}-\hat{\boldsymbol{\mu}}_k^{\mathrm{o}}$, $\boldsymbol{r}_k^{\mathrm{I}}=\boldsymbol{y}_k^{\mathrm{I}}-\hat{\boldsymbol{\mu}}_k^{\mathrm{M}}$, $\hat{V}_k^{(\mathrm{oM})}$ is the sub-matrix of $\hat{V}_k$ corresponding to the observed rows and missing columns of subject k, and $\hat{V}_k^{(\mathrm{MM})}$ the sub-matrix corresponding to missing rows and missing columns of k. And the double summation term could be safely dropped for simple random sampling and stratified random sampling.

4.3.3 Estimation of Cross Term Variance Component

To get an estimator of the mixed term $C_{\xi(I\pi)}$ (and $C_{\text{Imp-Sam}}$), we can follow the same idea in the previous two sections. We estimate $E_\xi(\boldsymbol{e}_i^{\mathrm{I}}\boldsymbol{e}_j^{\mathrm{o}\prime})$ by $\boldsymbol{r}_i^{\mathrm{I}}\boldsymbol{r}_j^{\mathrm{o}\prime}$, where $\boldsymbol{r}_i^{\mathrm{I}}=\boldsymbol{y}_i^{\mathrm{I}}-\hat{\boldsymbol{\mu}}_i^{\mathrm{M}}$ and $\boldsymbol{r}_j^{\mathrm{o}}=\boldsymbol{y}_j^{\mathrm{o}}-\hat{\boldsymbol{\mu}}_j^{\mathrm{o}}$; and we estimate $E_\xi(\boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{M}\prime})$ and $E_\xi(\boldsymbol{e}_i^{\mathrm{M}}\boldsymbol{e}_i^{\mathrm{M}\prime})$ by $\hat{V}_i^{(\mathrm{oM})}$ and $\hat{V}_i^{(\mathrm{MM})}$, respectively. Therefore, our estimator of $C_{\text{Imp-Sam}}$ and $C_{\xi(I\pi)}$ is

$$\begin{aligned}\hat{C}_{\text{Imp-Sam}}&=\hat{C}_{\xi(I\pi)}\\&=\big[\hat{H}(\hat{\boldsymbol{\beta}})\big]^{-1}\times\\&\quad\Bigg[\sum_{i,j\in s}\sum_{i\neq j}w_iw_j\frac{\partial\hat{\boldsymbol{\mu}}_i'}{\partial\hat{\boldsymbol{\beta}}}\hat{V}_i^{-1}\begin{pmatrix}\mathbf{0}&\mathbf{0}\\ \boldsymbol{r}_i^{\mathrm{I}}\boldsymbol{r}_j^{\mathrm{o}\prime}&\mathbf{0}\end{pmatrix}\hat{V}_j^{-1}\frac{\partial\hat{\boldsymbol{\mu}}_j}{\partial\hat{\boldsymbol{\beta}}}-\sum_s w_i^2\frac{\partial\hat{\boldsymbol{\mu}}_i'}{\partial\hat{\boldsymbol{\beta}}}\hat{V}_i^{-1}\begin{pmatrix}\mathbf{0}&\mathbf{0}\\ \hat{V}_i^{(\mathrm{oM})\prime}&\hat{V}_i^{(\mathrm{MM})}\end{pmatrix}\hat{V}_i^{-1}\frac{\partial\hat{\boldsymbol{\mu}}_i}{\partial\hat{\boldsymbol{\beta}}}\Bigg]\\&\quad\times\big[\hat{H}(\hat{\boldsymbol{\beta}})\big]^{-1}.\end{aligned}\quad(4.24)$$

4.4 An Alternative Approach for Variance Estimation

An alternative and simpler way of estimating the total variance of $\hat{\boldsymbol{\beta}}$ consists of expressing its total error around $\boldsymbol{\beta}$ directly, rather than decomposing it in the three terms, imputation, sampling, and model errors. Since $\hat{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}$, we can use Taylor series expansions to get

$$\begin{aligned}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}&=-\Big[E_{\pi I}\Big(\frac{\partial U_n^*(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\Big)\Big]^{-1}U_n^*(\boldsymbol{\beta})+o_p(1/\sqrt{r})\\&=[H(\boldsymbol{\beta})]^{-1}U_n^*(\boldsymbol{\beta})+o_p(1/\sqrt{r}),\end{aligned}$$

so that

$$\begin{aligned}V_{\text{Tot}}=&E_{\xi\pi RI}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\\=&E_{\xi R}[\text{Var}_{\pi I}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]\end{aligned}$$

$$= E_{\xi R}\{[H(\boldsymbol{\beta})]^{-1}\mathrm{Var}_{\pi I}[U_n^*(\boldsymbol{\beta})][H(\boldsymbol{\beta})]^{-1}\} + o(1/r),$$

where the "meat" is

$$\mathrm{Var}_{\pi I}[U_n^*(\boldsymbol{\beta})] = \mathrm{Var}_{\pi} E_I[U_n^*(\boldsymbol{\beta})] + E_{\pi}\mathrm{Var}_I[U_n^*(\boldsymbol{\beta})]. \tag{4.25}$$

The interchange of R and π in the expectation above can be justified in two ways. For one thing, if the individual missing probability is independent of other units selected in the sample, then R and π are clearly exchangeable. The other way to justify this is to note that under the current formulation of imputation classes, responses are missing completely at random within each class, and therefore the missing mechanism R can be ignored.

The two pieces in 4.25 can be called "sampling variance component" and "imputation variance component". The inner part of the sampling variance component is:

$$\begin{aligned} E_I[U_n^*(\boldsymbol{\beta})] &= E_I\Big[\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \begin{pmatrix} \boldsymbol{y}_i^{\mathrm{o}} - \boldsymbol{\mu}_i^{\mathrm{o}} \\ \boldsymbol{y}_i^{\mathrm{I}} - \boldsymbol{\mu}_i^{\mathrm{M}} \end{pmatrix}\Big] \\ &= \sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \begin{pmatrix} \boldsymbol{y}_i^{\mathrm{o}} - \boldsymbol{\mu}_i^{\mathrm{o}} \\ \bar{y}_{\tau r} - \boldsymbol{\mu}_i^{\mathrm{M}} \end{pmatrix}. \end{aligned} \tag{4.26}$$

We then must calculate the design variance of this quantity. If we called

$$z_{\tau i} = \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \begin{pmatrix} \boldsymbol{y}_i^{\mathrm{o}} - \boldsymbol{\mu}_i^{\mathrm{o}} \\ \bar{y}_{\tau r} - \boldsymbol{\mu}_i^{\mathrm{M}} \end{pmatrix},$$

then $E_I[U_n^*(\boldsymbol{\beta})] = \sum_s w_i z_{\tau i}$ would look like a Horvitz-Thompson estimator; for which we can use usual design techniques to calculate its variance and an estimate of it. Let $V_{\tau\text{naïve}}$ be such variance and $\hat{V}_{\tau\text{naïve}}$ be the estimator.

Since there are some values in some of the $z_{\tau i}$'s that are not actual observed quantities, but sample-based means, 4.26 is not exactly a H-T estimator. $V_{\tau\text{naïve}}$ tends to be too small compared to the true design variance of $E_I[U_n^*(\boldsymbol{\beta})] = \sum_s w_i z_{\tau i}$. In simple cases it can be seen that $V_{\tau\text{naïve}} = (r/n)^2\mathrm{Var}_{\pi}[\sum_s w_i z_{\tau i}]$. If we assume that $\hat{V}_{\tau\text{naïve}}$ is an unbiased estimator of $V_{\tau\text{naïve}}$, then $(n/r)^2\hat{V}_{\tau\text{naïve}}$ would be an unbiased estimator of (the "meat" of) the sampling variance component, $\mathrm{Var}_{\pi}E_I[U_n^*(\boldsymbol{\beta})]$.

The second piece of 4.25 corresponds to the imputation variance component, and can be easily computed as follows. The inner part is:

$$\begin{aligned} \mathrm{Var}_I[U_n^*(\boldsymbol{\beta})] &= \mathrm{Var}_I\Big[\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \begin{pmatrix} \boldsymbol{y}_i^{\mathrm{o}} - \boldsymbol{\mu}_i^{\mathrm{o}} \\ \boldsymbol{y}_i^{\mathrm{I}} - \boldsymbol{\mu}_i^{\mathrm{M}} \end{pmatrix}\Big] \\ &= \sum_s w_i^2 \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(s_{\tau r}^2) \end{pmatrix} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}; \end{aligned} \tag{4.27}$$

which can be directly estimated by plugging in the corresponding estimated values of $\boldsymbol{\beta}$ and association parameter(s). Obviously such estimator also estimates (the "meat"of) the imputation variance component, $E_{\pi}\mathrm{Var}_I[U_n^*(\boldsymbol{\beta})]$.

4.5 Proofs

4.5.1 Proof of Theorem 4.2: Consistency of PGEE Estimators with Hot-deck Imputation

Lemma 4.4. *Suppose that conditions in Lemma 3.2 hold and that Y_i^* is the vector Y_i with missing values imputed by the hot-deck method (weighted or unweighted), then, as $r,n,N \to \infty$,*

$$\sup_{\boldsymbol{\beta}\in\Theta}\left|\left|\frac{1}{N}s_n^*(\boldsymbol{\beta})-\Delta_N^*(\boldsymbol{\beta})\right|\right| \xrightarrow{p} 0,$$

where $s_n^*(\boldsymbol{\beta})=\sum_s w_i\psi_i(Y_i^*,\boldsymbol{\beta})$ and $\Delta_N^*(\boldsymbol{\beta})=E_{\xi\pi RI}[N^{-1}s_n^*(\boldsymbol{\beta})]=N^{-1}E_{\xi\pi RI}\sum_s w_i\psi_i(Y_i^*,\boldsymbol{\beta})$.

Proof of Lemma 4.4:
The proof follows the same lines as the proof of Lemma 3.2, with modifications over terms involving Y_i^*. Expression 3.7 remains unchanged; which implies that $P(h_j(Y_j)I(\|Y_j\|>c)>\varepsilon)\le E_\xi[h_j(Y_j)I(\|Y_j\|>c)]/\varepsilon = O(c^{-1/2})$, independent of j. Now, for any $\mathscr{O}\subset\Theta$:

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\mathscr{O}}\left|\psi_i(Y_i^*,\boldsymbol{\beta})\right|I(\|Y_i^*\|>c) &\le \sup_{\boldsymbol{\beta}\in\Theta}\left|\psi_i(Y_i^*,\boldsymbol{\beta})\right|I(\|Y_i^*\|>c)\\ &\le \sup_j\{\sup_{\boldsymbol{\beta}\in\Theta}\left|\psi_j(Y_j,\boldsymbol{\beta})\right|I(\|Y_j\|>c)\}\\ &= \sup_j\{h_j(Y_j)I(\|Y_j\|>c)\},\end{aligned}$$

then,

$$\begin{aligned}\left|\frac{1}{N}\sum_{i\in s}w_i\sup_{\boldsymbol{\beta}\in\mathscr{O}}\psi_i(Y_i^*,\boldsymbol{\beta})I(\|Y_i^*\|>c)\right| &\le \frac{1}{N}\sum_{i\in s}w_i\sup_{\boldsymbol{\beta}\in\mathscr{O}}|\psi_i(Y_i^*,\boldsymbol{\beta})|I(\|Y_i^*\|>c)\\ &\le \left[\frac{1}{N}\sum_{i\in s}w_i\right]\sup_j\{h_j(Y_j)I(\|Y_j\|>c)\}\\ &= O_p(1)O_p(c^{-1/2})=O_p(c^{-1/2}).\end{aligned}$$

This means that for any $\mathscr{O}\subset\Theta$, as $c\to\infty$,

$$P_{\xi\pi RI}\left(\frac{1}{N}\sum_s w_i\sup_{\boldsymbol{\beta}\in\mathscr{O}}\psi_i(Y_i^*,\boldsymbol{\beta})I(\|Y_i^*\|>c)-\frac{1}{N}\sum_s w_i\inf_{\boldsymbol{\beta}\in\mathscr{O}}\psi_i(Y_i^*,\boldsymbol{\beta})I(\|Y_i^*\|>c)>\varepsilon\right)\to 0. \tag{4.28}$$

Now, by the equicontinuity of $\{g_i(\boldsymbol{\beta})=\psi_i(Y_i,\boldsymbol{\beta})\}$ over i and Θ, given any $\varepsilon>0$, there exists a δ, such that for any open ball $\mathscr{O}$ with radius less than δ, if $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2\in\mathscr{O}$, we have:

$$\left|\psi_i(Y,\boldsymbol{\beta}_1)I(\|Y\|\le c)-\psi_i(Y,\boldsymbol{\beta}_2)I(\|Y\|\le c)\right|\le\varepsilon/4,$$

for any Y. Hence, for the same $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$,

$$\left|\psi_i(Y^*,\boldsymbol{\beta}_1)I(\|Y^*\|\le c)-\psi_i(Y^*,\boldsymbol{\beta}_2)I(\|Y^*\|\le c)\right|\le\varepsilon/4.$$

From here we continue as in section 3.3 to get that for any open ball $\mathscr{O}$ with radius less than δ, as $r \to \infty$,

$$P_{\xi\pi RI}\left(\frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^*,\boldsymbol{\beta}) - \frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^*,\boldsymbol{\beta}) > \frac{\varepsilon}{2}\right) \to 0. \tag{4.29}$$

Expression 3.13 remains unchanged. On the other hand, if the donors are selected with probabilities proportional to τ_i, $E_I[u_i(Y_i^{\mathrm{I}})] = \sum_r \tau_i u_i(Y_i)/\sum_r \tau_i$ for any suitable function $u_i(\cdot)$. If donors are selected by weighted hot-deck ($\tau_i = w_i$),

$$\frac{\sum_r w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})}{\sum_r w_i} = E_I[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta})],$$

and if donors are selected by unweighted hot-deck ($\tau_i \equiv 1$),

$$\begin{aligned}
\frac{\sum_r w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})}{\sum_r w_i} &= \frac{\sum_{i=1}^{N_r} \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})}{N_r} + O_p(1/\sqrt{r}) \\
&= E_\xi[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})] + O_p(1/\sqrt{r}) \\
&= \frac{\sum_r \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})}{r} + O_p(1/\sqrt{r}) \\
&= E_I[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta})] + O_p(1/\sqrt{r});
\end{aligned}$$

so that, in any case, by the MAR assumption,

$$\begin{aligned}
\frac{\sum_m w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})}{\sum_m w_i} &= \frac{\sum_r w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})}{\sum_r w_i} + O_p(1/\sqrt{r}) \\
&= E_I[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta})] + O_p(1/\sqrt{r}).
\end{aligned}$$

Then we get

$$\begin{aligned}
&\frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^*,\boldsymbol{\beta}) - \frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta}) \\
&\quad= \frac{1}{N}\sum_m w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta}) - \frac{1}{N}\sum_m w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta}) \\
&\quad= \frac{1}{N}\sum_m w_i \Big[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta}) - \frac{\sum_m w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})}{\sum_m w_i}\Big] \\
&\quad= \frac{1}{N}\sum_m w_i \Big[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta}) - E_I[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta})]\Big] + O_p(1/\sqrt{r}) \\
&\quad= \frac{1}{\sqrt{N}}\frac{\sqrt{N_m}}{\sqrt{N}}\frac{1}{\sqrt{N_m}}\sum_{i=1}^{N_m}\Big[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta}) - E_I[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta})]\Big] + O_p(1/\sqrt{r}) \\
&\quad= O_p(1/\sqrt{N})O_p(1)O_p(1) + O_p(1/\sqrt{r}) = O_p(1/\sqrt{r}),
\end{aligned}$$

where the second to last line is due to the fact that, because of Chebyshev's inequality, $N_m^{-1/2} \times \sum_{i=1}^{N_m} \{\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta}) - E_I[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^{\mathrm{I}},\boldsymbol{\beta})\} = O_p(1)$. This implies that

$$P_{\xi\pi RI}\left(\left|\left|\frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^*,\boldsymbol{\beta}) - \frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})\right|\right| > \frac{\varepsilon}{2}\right) \to 0. \tag{4.30}$$

Therefore, by 4.30 and 3.13,

$$P_{\xi\pi RI}\left(\left|\left|\frac{1}{N}\sum_s w_i \inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^*,\boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^{N} E_\xi\left[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})\right]\right|\right| > \varepsilon\right) \to 0. \tag{4.31}$$

Based on 4.29 and 4.31, we can show, similarly to expression 3.14, that

$$P_{\xi\pi RI}\left(\frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^*,\boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^{N} E_\xi\left[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})\right] > \varepsilon\right) \to 0. \tag{4.32}$$

Now let $H_n(\boldsymbol{\beta}) = \frac{1}{N}\sum_s w_i\psi_i(Y_i^*,\boldsymbol{\beta}) - E_{\xi\pi RI}[\frac{1}{N}\sum_s w_i\psi_i(Y_i^*,\boldsymbol{\beta})]$; then

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in\mathscr{O}} H_n(\boldsymbol{\beta}) &= \sup_{\boldsymbol{\beta}\in\mathscr{O}}\left\{\frac{1}{N}\sum_s w_i\psi_i(Y_i^*,\boldsymbol{\beta}) - E_{\pi RI}\left[\frac{1}{N}\sum_s w_i E_\xi[\psi_i(Y_i^*,\boldsymbol{\beta})]\right]\right\} \\
&= \sup_{\boldsymbol{\beta}\in\mathscr{O}}\left\{\frac{1}{N}\sum_s w_i\psi_i(Y_i^*,\boldsymbol{\beta}) - E_{\pi RI}\left[\frac{1}{N}\sum_s w_i E_\xi[\psi_i(Y_i,\boldsymbol{\beta})]\right]\right\} \qquad (4.33)\\
&= \sup_{\boldsymbol{\beta}\in\mathscr{O}}\left\{\frac{1}{N}\sum_s w_i\psi_i(Y_i^*,\boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^{N} E_\xi[\psi_i(Y_i,\boldsymbol{\beta})]\right\} \\
&\le \frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^*,\boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^{N} E_\xi\left[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})\right],
\end{aligned}$$

where 4.33 depends on $E_\xi[\psi_i(Y_i^*,\boldsymbol{\beta})] = E_\xi[\psi_i(Y_i,\boldsymbol{\beta})]$ holding[2], and the last line can be derived exactly as in section 3.3; so that, by 4.32,

$$\begin{aligned}
P_{\xi\pi RI}\left(\sup_{\boldsymbol{\beta}\in\mathscr{O}} H_n(\boldsymbol{\beta}) > \varepsilon\right) &\le P_{\xi\pi RI}\left(\frac{1}{N}\sum_s w_i \sup_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i^*,\boldsymbol{\beta}) - \frac{1}{N}\sum_{i=1}^{N} E_\xi\left[\inf_{\boldsymbol{\beta}\in\mathscr{O}} \psi_i(Y_i,\boldsymbol{\beta})\right] > \varepsilon\right) \\
&\to 0.
\end{aligned} \tag{4.34}$$

Similarly, it can be shown that

$$P_{\xi\pi RI}\left(\inf_{\boldsymbol{\beta}\in\mathscr{O}} H_n(\boldsymbol{\beta}) < -\varepsilon\right) \to 0. \tag{4.35}$$

[2] See the comments after the Proof of Theorem 4.2.

Using 4.34 and 4.35, we can get, as in the proof of Lemma 3.2, $P_{\xi\pi}\big(\sup_{\boldsymbol{\beta}\in\Theta}|H_n(\boldsymbol{\beta})| > \varepsilon\big) \to 0$. And we have the desired result:

$$\begin{aligned}\sup_{\boldsymbol{\beta}\in\Theta}|H_n(\boldsymbol{\beta})| &= \sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{N}\sum_s w_i\psi_i(Y_i^*,\boldsymbol{\beta}) - E_{\xi\pi RI}\left[\frac{1}{N}\sum_s w_i\psi_i(Y_i^*,\boldsymbol{\beta})\right]\right| \\ &= \sup_{\boldsymbol{\beta}\in\Theta}\left|\frac{1}{N}s_n^*(\boldsymbol{\beta}) - \Delta_N^*(\boldsymbol{\beta})\right| \xrightarrow{p} 0.\end{aligned}$$

□

Proof of Theorem 4.2: Without loss of generality we consider cases in which $E_{\xi I}[\psi_i(Y_i^*,\boldsymbol{\beta})] = E_{\xi I}[\psi_i(Y_i,\boldsymbol{\beta})]$. Once we notice that, for any $\boldsymbol{\beta}\in\Theta$,

$$\begin{aligned}\Delta_N^*(\boldsymbol{\beta}) =& E_{\pi R}\left[\frac{1}{N}\sum_s w_i E_{\xi I}[\psi_i(Y_i^*,\boldsymbol{\beta})]\right] \\ =& E_{\pi R}\left[\frac{1}{N}\sum_s w_i E_{\xi I}[\psi_i(Y_i,\boldsymbol{\beta})]\right] \\ =& \Delta_N(\boldsymbol{\beta}),\end{aligned}$$

the proof follows the same lines as the Proof of Theorem 3.1, but making use of Lemma 4.4 this time. □

COMMENTS:

1. GEE is semi-parametric. However, in practical situations $Y_1, Y_2, \ldots, Y_n$ can usually be viewed as iid observations within each imputation class, for which the set of covariates has identical values. So that, within imputation classes, $E_\xi[\psi_i(Y_i^*,\boldsymbol{\beta})] = E_\xi[\psi_i(Y_i,\boldsymbol{\beta})]$.
2. To a less restricted situation in this thesis, the $\psi_i(Y_i,\boldsymbol{\beta})$ is indeed linear in Y_i. Within each imputation class we have $E_\xi[Y_i^*] = E_\xi[Y_i]$, and hence $E_\xi[\psi_i(Y_i^*,\boldsymbol{\beta})] = E_\xi[\psi_i(Y_i,\boldsymbol{\beta})]$.
3. Under both scenarios, 1 or 2, we have $\Delta_N^*(\boldsymbol{\beta}) = \Delta_N(\boldsymbol{\beta})$. Nonetheless, even if this equality does not strictly hold, the proof remains valid as long as $\Delta_N^*(\boldsymbol{\beta}) = \Delta_N(\boldsymbol{\beta}) + o(1)$.

4.5.2 Proof of Theorem 4.3: Variance Decomposition and Estimation

Variance Components

Proof of Theorem 4.3.
We begin with $V_{I(I)}$, the variance due to imputation is $V_{\text{Imp}} = E_{\xi\pi R}V_{I(I)}$. Since $\hat{\boldsymbol{\beta}}$ solves 4.12,

$$\mathbf{0} = U_n^*(\hat{\boldsymbol{\beta}})$$

$$
\begin{aligned}
&= U_n^*(\hat{\boldsymbol{\beta}}_n) + \frac{\partial U_n^*(\hat{\boldsymbol{\beta}}_n)}{\partial \hat{\boldsymbol{\beta}}_n}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n) + o_p(1/\sqrt{r}) \\
&= U_n^*(\hat{\boldsymbol{\beta}}_n) + E_I\Big(\frac{\partial U_n^*(\hat{\boldsymbol{\beta}}_n)}{\partial \hat{\boldsymbol{\beta}}_n}\Big)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n) + \Big[\frac{\partial U_n^*(\hat{\boldsymbol{\beta}}_n)}{\partial \hat{\boldsymbol{\beta}}_n} - E_I\Big(\frac{\partial U_n^*(\hat{\boldsymbol{\beta}}_n)}{\partial \hat{\boldsymbol{\beta}}_n}\Big)\Big](\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n) + o_p(1/\sqrt{r}) \\
&= U_n^*(\hat{\boldsymbol{\beta}}_n) + E_I\Big(\frac{\partial U_n^*(\hat{\boldsymbol{\beta}}_n)}{\partial \hat{\boldsymbol{\beta}}_n}\Big)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n) + o_p(1/\sqrt{r}).
\end{aligned}
$$

Therefore,

$$
\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n = -\Big[E_I\Big(\frac{\partial U_n^*(\hat{\boldsymbol{\beta}}_n)}{\partial \hat{\boldsymbol{\beta}}_n}\Big)\Big]^{-1} U_n^*(\hat{\boldsymbol{\beta}}_n) + o_p(1/\sqrt{r}),
$$

so that

$$
\begin{aligned}
V_{I(I)} &= E_I(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)' \\
&= \Big[E_I\Big(\frac{\partial U_n^*(\hat{\boldsymbol{\beta}}_n)}{\partial \hat{\boldsymbol{\beta}}_n}\Big)\Big]^{-1} E_I[U_n^*(\hat{\boldsymbol{\beta}}_n)(U_n^*(\hat{\boldsymbol{\beta}}_n))']\Big[E_I\Big(\frac{\partial U_n^*(\hat{\boldsymbol{\beta}}_n)}{\partial \hat{\boldsymbol{\beta}}_n}\Big)\Big]^{-1'} + o_p(1/r). \qquad (4.36)
\end{aligned}
$$

Similarly we can get a corresponding expression for V_π,

$$
\hat{\boldsymbol{\beta}}_n - B = -\Big[E_\pi\Big(\frac{\partial U_n(B)}{\partial B}\Big)\Big]^{-1} U_n(B) + o_p(1/\sqrt{n}),
$$

and

$$
V_\pi = \mathrm{Var}_\pi(\hat{\boldsymbol{\beta}}_n - B) = \Big[E_\pi\Big(\frac{\partial U_n(B)}{\partial B}\Big)\Big]^{-1} \mathrm{Var}_\pi[U_n(B)]\Big[E_\pi\Big(\frac{\partial U_n(B)}{\partial B}\Big)\Big]^{-1'} + o_p(1/n). \quad (4.37)
$$

Similarly we can get the following:

$$
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = -\Big[E_\xi\Big(\frac{\partial U_n^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big)\Big]^{-1} U_n^*(\boldsymbol{\beta}) + o_p(1/\sqrt{r}),
$$

$$
\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta} = -\Big[E_\xi\Big(\frac{\partial U_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big)\Big]^{-1} U_n(\boldsymbol{\beta}) + o_p(1/\sqrt{n}).
$$

And then an expression for $C_{\xi(I\pi)}$ is:

$$
\begin{aligned}
C_{\xi(I\pi)} &= E_\xi(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\beta}}_n - B)' \\
&= E_\xi[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_n)][(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})' + (\boldsymbol{\beta} - B)'] \\
&= E_\xi(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})' - E_\xi(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})' + o_p(1/r) \\
&= \Big[E_\xi\Big(\frac{\partial U_n^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big)\Big]^{-1} E_\xi[U_n^*(\boldsymbol{\beta})U_n'(\boldsymbol{\beta})]\Big[E_\xi\Big(\frac{\partial U_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big)\Big]^{-1'} \\
&\quad - \Big[E_\xi\Big(\frac{\partial U_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big)\Big]^{-1} E_\xi[U_n(\boldsymbol{\beta})U_n'(\boldsymbol{\beta})]\Big[E_\xi\Big(\frac{\partial U_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big)\Big]^{-1'} + o_p(1/r). \qquad (4.38)
\end{aligned}
$$

Furthermore, we should get expressions for each of the terms in $V_{I(I)}$, V_π, and $C_{\xi(I\pi)}$. We begin with $\partial U_n^*(\hat{\boldsymbol{\beta}}_n)/\partial\hat{\boldsymbol{\beta}}_n$. Since

$$\begin{aligned}\frac{1}{N}\sum_s w_i \underbrace{\frac{\partial}{\partial\boldsymbol{\beta}}\Big(\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\Big)}_{A_i}(\boldsymbol{y}_i^*-\boldsymbol{\mu}_i) &= \frac{1}{N}\sum_s w_i A_i\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}}\\ \boldsymbol{y}_i^{\mathrm{I}}\end{pmatrix} - \frac{1}{N}\sum_s w_i A_i\begin{pmatrix}\boldsymbol{\mu}_i^{\mathrm{o}}\\ \boldsymbol{\mu}_i^{\mathrm{M}}\end{pmatrix}\\ &= \frac{1}{N}\sum_s w_i A_i\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}}\\ \mathbf{0}\end{pmatrix} - \frac{1}{N}\sum_s w_i A_i E_\xi\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}}\\ \mathbf{0}\end{pmatrix} + \frac{1}{N}\sum_s w_i A_i\begin{pmatrix}\mathbf{0}\\ \boldsymbol{y}_i^{\mathrm{I}}\end{pmatrix} - \frac{1}{N}\sum_s w_i A_i E_\xi\begin{pmatrix}\mathbf{0}\\ \boldsymbol{y}_i^{\mathrm{I}}\end{pmatrix}\\ &= O_p(1/\sqrt{r}),\end{aligned}$$

then,

$$\begin{aligned}\frac{\partial U_n^*(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} &= \frac{\partial}{\partial\boldsymbol{\beta}}\Big[\frac{1}{N}\sum_s w_i\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}(\boldsymbol{y}_i^*-\boldsymbol{\mu}_i)\Big]\\ &= \frac{1}{N}\sum_s w_i\frac{\partial}{\partial\boldsymbol{\beta}}\Big(\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\Big)(\boldsymbol{y}_i^*-\boldsymbol{\mu}_i) - \frac{1}{N}\sum_s w_i\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}\\ &= -\frac{1}{N}\sum_s w_i\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}} + O_p(1/\sqrt{r});\end{aligned} \tag{4.39}$$

but since $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}$, then also $\partial U_n^*(\hat{\boldsymbol{\beta}}_n)/\partial\hat{\boldsymbol{\beta}}_n = -N^{-1}\sum_s w_i\, {}^{\partial\boldsymbol{\mu}_i'}\!/_{\partial\hat{\boldsymbol{\beta}}_n} V_i^{-1}\, {}^{\partial\boldsymbol{\mu}_i}\!/_{\partial\hat{\boldsymbol{\beta}}_n} + O_p(1/\sqrt{r})$.

Likewise we have,

$$\frac{\partial U_n(B)}{\partial B} = -\frac{1}{N}\sum_s w_i\frac{\partial\boldsymbol{\mu}_i'}{\partial B}V_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial B} + O_p(1/\sqrt{n}). \tag{4.40}$$

In summary, we obtain the following four expressions,

$$\begin{aligned}E_I\big(\partial U_n^*(\hat{\boldsymbol{\beta}}_n)/\partial\hat{\boldsymbol{\beta}}_n\big) &= -\hat{H}(\hat{\boldsymbol{\beta}}_n) + O_p(1/\sqrt{r}), \quad E_\xi\big(\partial U_n^*(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\big) = -\hat{H}(\boldsymbol{\beta}) + O_p(1/\sqrt{r}),\\ E_\pi\big(\partial U_n(B)/\partial B\big) &= -H(B) + O_p(1/\sqrt{n}), \quad E_\xi\big(\partial U_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\big) = -\hat{H}(\boldsymbol{\beta}) + O_p(1/\sqrt{n}).\end{aligned}$$

These terms account for the "bread" parts of $V_{I(I)}$, V_π, and $C_{\xi(I\pi)}$. Now we turn to the inner pieces (the "meat" terms) in each of them. These are the variances (or second moments) of the corresponding estimating functions.

First recall that under unweighted hot-deck imputation in each cycle/cell, $y_i^{\mathrm{I}} = y_j$ if respondent j is the donor for nonrespondent i, which occurs with probability $1/r$. Then, $E_I(y_i^{\mathrm{I}}) = \bar{y}_r$, the mean of respondents; $E_I\big((y_i^{\mathrm{I}})^2\big) = \overline{y_r^2}$, the mean of the square of respondents; so that $\mathrm{Var}_I(y_i^{\mathrm{I}}) = \overline{y_r^2} - \bar{y}_r^2$. On the other hand, if we use weighted hot-deck, $y_i^{\mathrm{I}} = y_j$ with probability $w_j/\sum_r w_j$. Then $E_I(y_i^{\mathrm{I}}) = \sum_r w_j y_j/\sum_r w_j$ (the weighted mean of respondents), and $E_I\big((y_i^{\mathrm{I}})^2\big) = \sum_r w_j y_j^2/\sum_r w_j$ (the weighted mean of the squared responses); and $\mathrm{Var}_I(y_i^{\mathrm{I}}) = (\sum_r w_j y_j^2/\sum_r w_j) - (\sum_r w_j y_j/\sum_r w_j)^2$.

We can summarize this by writing $E_I(y_i^{\mathrm{I}}) = \sum_r \tau_j y_j/\sum_r \tau_j = \bar{y}_{\tau r}$ and

$$\mathrm{Var}_I(y_i^{\mathrm{I}}) = \frac{\sum_r \tau_j y_j^2}{\sum_r \tau_j} - \Big(\frac{\sum_r \tau_j y_j}{\sum_r \tau_j}\Big)^2 = s_{\tau r}^2;$$

where $\tau_i \equiv 1$ under unweighted hot-deck and $\tau_i = w_i$ under weighted hot-deck. So that,

$$\begin{aligned}
&E_I(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)' \\
&\quad = \mathrm{Var}_I(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i) + E_I(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)E_I(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)' = \mathrm{Var}_I(\boldsymbol{y}_i^*) + (E_I\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)(E_I\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)' \\
&\quad = \mathrm{Var}_I\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} \\ \boldsymbol{y}_i^{\mathrm{I}}\end{pmatrix} + \left[E_I\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} \\ \boldsymbol{y}_i^{\mathrm{I}}\end{pmatrix} - \boldsymbol{\mu}_i\right]\left[E_I\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} \\ \boldsymbol{y}_i^{\mathrm{I}}\end{pmatrix} - \boldsymbol{\mu}_i\right]' \\
&\quad = \begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(s^2_{\tau r})\end{pmatrix} + \left[\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} \\ \bar{\boldsymbol{y}}_{\tau r}\end{pmatrix} - \boldsymbol{\mu}_i\right]\left[\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} \\ \bar{\boldsymbol{y}}_{\tau r}\end{pmatrix} - \boldsymbol{\mu}_i\right]';
\end{aligned}$$

and, for $i \neq j$,

$$E_I(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)(\boldsymbol{y}_j^* - \boldsymbol{\mu}_j)' = \left[E_I\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} \\ \boldsymbol{y}_i^{\mathrm{I}}\end{pmatrix} - \boldsymbol{\mu}_i\right]\left[E_I\begin{pmatrix}\boldsymbol{y}_j^{\mathrm{o}} \\ \boldsymbol{y}_j^{\mathrm{I}}\end{pmatrix} - \boldsymbol{\mu}_j\right]' = \left[\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} \\ \bar{\boldsymbol{y}}_{\tau r}\end{pmatrix} - \boldsymbol{\mu}_i\right]\left[\begin{pmatrix}\boldsymbol{y}_j^{\mathrm{o}} \\ \bar{\boldsymbol{y}}_{\tau r}\end{pmatrix} - \boldsymbol{\mu}_j\right]'.$$

Therefore, for the term $E_I[U_n^*(\hat{\boldsymbol{\beta}}_n)(U_n^*(\hat{\boldsymbol{\beta}}_n))']$ in expression 4.36 for $V_{I(I)}$ we have:

$$\begin{aligned}
&E_I[U_n^*(\hat{\boldsymbol{\beta}}_n)(U_n^*(\hat{\boldsymbol{\beta}}_n))'] \\
&\quad = E_I\Big[\frac{1}{N}\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1}(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i) \cdot \frac{1}{N}\sum_s w_i(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)' V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \hat{\boldsymbol{\beta}}_n}\Big] \\
&\quad = \frac{1}{N^2}E_I\Big[\sum_s w_i^2 \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1}(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)' V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \hat{\boldsymbol{\beta}}_n} \\
&\qquad + \sum_{i,j\in s}\sum_{i\neq j} w_i w_j \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1}(\boldsymbol{y}_i^* - \boldsymbol{\mu}_i)(\boldsymbol{y}_j^* - \boldsymbol{\mu}_j)' V_j^{-1}\frac{\partial \boldsymbol{\mu}_j}{\partial \hat{\boldsymbol{\beta}}_n}\Big] \\
&\quad = \frac{1}{N^2}\Big\{\sum_s w_i^2 \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(s^2_{\tau r})\end{pmatrix} V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \hat{\boldsymbol{\beta}}_n} + \sum_s w_i^2 \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1}\left[\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} \\ \bar{\boldsymbol{y}}_{\tau r}\end{pmatrix} - \boldsymbol{\mu}_i\right]^{\otimes 2} V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \hat{\boldsymbol{\beta}}_n} \\
&\qquad + \sum_{i,j\in s}\sum_{i\neq j} w_i w_j \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1}\left[\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} \\ \bar{\boldsymbol{y}}_{\tau r}\end{pmatrix} - \boldsymbol{\mu}_i\right]\left[\begin{pmatrix}\boldsymbol{y}_j^{\mathrm{o}} \\ \bar{\boldsymbol{y}}_{\tau r}\end{pmatrix} - \boldsymbol{\mu}_j\right]' V_j^{-1}\frac{\partial \boldsymbol{\mu}_j}{\partial \hat{\boldsymbol{\beta}}_n}\Big\} \\
&\quad = \frac{1}{N^2}\Big\{\sum_s w_i^2 \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(s^2_{\tau r})\end{pmatrix} V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \hat{\boldsymbol{\beta}}_n} \\
&\qquad + \sum_{i\in s}\sum_{j\in s} w_i w_j \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1}\left[\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} \\ \bar{\boldsymbol{y}}_{\tau r}\end{pmatrix} - \boldsymbol{\mu}_i\right]\left[\begin{pmatrix}\boldsymbol{y}_j^{\mathrm{o}} \\ \bar{\boldsymbol{y}}_{\tau r}\end{pmatrix} - \boldsymbol{\mu}_j\right]' V_j^{-1}\frac{\partial \boldsymbol{\mu}_j}{\partial \hat{\boldsymbol{\beta}}_n}\Big\} \\
&\quad = \frac{1}{N^2}\Big\{\sum_s w_i^2 \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(s^2_{\tau r})\end{pmatrix} V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \hat{\boldsymbol{\beta}}_n} + \Big[\sum_{i\in s} w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial \hat{\boldsymbol{\beta}}_n} V_i^{-1}\begin{pmatrix}\boldsymbol{y}_i^{\mathrm{o}} - \boldsymbol{\mu}_i^{\mathrm{o}} \\ \bar{\boldsymbol{y}}_{\tau r} - \boldsymbol{\mu}_i^{\mathrm{M}}\end{pmatrix}\Big]^{\otimes 2}\Big\},
\end{aligned} \tag{4.41}$$

where $s^2_{\tau r}$ is the τ-weighted sample variance and $\bar{y}_{\tau r}$ the τ-weighted mean of respondents in the given cycle.

Since, conditioning on the model ξ, $U_n(B)$ is a Horvitz-Thompson estimator, we get the following for $\mathrm{Var}_\pi[U_n(B)]$ in expression 4.37 for V_π:

$$\mathrm{Var}_\pi[U_n(B)] = \mathrm{Var}_\pi\Big[\frac{1}{N}\sum_s w_i \frac{\partial \boldsymbol{\mu}_i'}{\partial B} V_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\Big]$$

$$
\begin{aligned}
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\Delta_{ij}}{\pi_i\pi_j}\Big[\frac{\partial \boldsymbol{\mu}_i'}{\partial B}V_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\Big]\Big[\frac{\partial \boldsymbol{\mu}_j'}{\partial B}V_j^{-1}(\boldsymbol{y}_j-\boldsymbol{\mu}_j)\Big]' \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\Delta_{ij}}{\pi_i\pi_j}\boldsymbol{z}_i\boldsymbol{z}_j'; \qquad (4.42)
\end{aligned}
$$

where π_{ij} is the joint probability of inclusion of elements i and j; see for example page 170 in Särndal et al. (1992).

With regard to $E_\xi[U_n^*(\boldsymbol{\beta})U_n'(\boldsymbol{\beta})]$ in expression 4.38 for $C_{\xi(I\pi)}$ we have the following. Since

$$
\begin{aligned}
U_n^*(\boldsymbol{\beta})U_n'(\boldsymbol{\beta}) &= \frac{1}{N}\sum_s w_i\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}(\boldsymbol{y}_i^*-\boldsymbol{\mu}_i)\cdot\frac{1}{N}\sum_s w_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)'V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{N^2}\sum_s w_i\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}\begin{pmatrix}\boldsymbol{e}_i^{\mathrm{o}}\\ \boldsymbol{e}_i^{\mathrm{I}}\end{pmatrix}\cdot\sum_s w_i\big(\boldsymbol{e}_i^{\mathrm{o}\prime},\ \boldsymbol{e}_i^{\mathrm{M}\prime}\big)V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{N^2}\Big\{\sum_s w_i^2\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}\begin{pmatrix}\boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{o}\prime} & \boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{M}\prime}\\ \boldsymbol{e}_i^{\mathrm{I}}\boldsymbol{e}_i^{\mathrm{o}\prime} & \boldsymbol{e}_i^{\mathrm{I}}\boldsymbol{e}_i^{\mathrm{M}\prime}\end{pmatrix}V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \\
&\qquad + \sum_{i,j\in s}\sum_{i\neq j} w_iw_j\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}\begin{pmatrix}\boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_j^{\mathrm{o}\prime} & \boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_j^{\mathrm{M}\prime}\\ \boldsymbol{e}_i^{\mathrm{I}}\boldsymbol{e}_j^{\mathrm{o}\prime} & \boldsymbol{e}_i^{\mathrm{I}}\boldsymbol{e}_j^{\mathrm{M}\prime}\end{pmatrix}V_j^{-1}\frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{\beta}}\Big\},
\end{aligned}
$$

then,

$$
\begin{aligned}
E_\xi[U_n^*(\boldsymbol{\beta})U_n'(\boldsymbol{\beta})] &= \frac{1}{N^2}\Big\{\sum_s w_i^2\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}\begin{pmatrix}E_\xi(\boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{o}\prime}) & E_\xi(\boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{M}\prime})\\ \mathbf{0} & \mathbf{0}\end{pmatrix}V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \\
&\qquad + \sum_{i,j\in s}\sum_{i\neq j} w_iw_j\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ E_\xi(\boldsymbol{e}_i^{\mathrm{I}}\boldsymbol{e}_j^{\mathrm{o}\prime}) & \mathbf{0}\end{pmatrix}V_j^{-1}\frac{\partial \boldsymbol{\mu}_j}{\partial \boldsymbol{\beta}}\Big\}. \qquad (4.43)
\end{aligned}
$$

And for $E_\xi[U_n(\boldsymbol{\beta})U_n'(\boldsymbol{\beta})]$, the following. Since

$$
\begin{aligned}
\mathrm{Var}_\xi(\boldsymbol{y}_i-\boldsymbol{\mu}_i) &= E_\xi(\boldsymbol{y}_i-\boldsymbol{\mu}_i)(\boldsymbol{y}_i-\boldsymbol{\mu}_i)' = E_\xi\Big[\begin{pmatrix}\boldsymbol{e}_i^{\mathrm{o}}\\ \boldsymbol{e}_i^{\mathrm{M}}\end{pmatrix}\cdot\big(\boldsymbol{e}_i^{\mathrm{o}\prime},\ \boldsymbol{e}_i^{\mathrm{M}\prime}\big)\Big] \\
&= E_\xi\begin{pmatrix}\boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{o}\prime} & \boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{M}\prime}\\ \boldsymbol{e}_i^{\mathrm{M}}\boldsymbol{e}_i^{\mathrm{o}\prime} & \boldsymbol{e}_i^{\mathrm{M}}\boldsymbol{e}_i^{\mathrm{M}\prime}\end{pmatrix} = \begin{pmatrix}E_\xi(\boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{o}\prime}) & E_\xi(\boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{M}\prime})\\ E_\xi(\boldsymbol{e}_i^{\mathrm{M}}\boldsymbol{e}_i^{\mathrm{o}\prime}) & E_\xi(\boldsymbol{e}_i^{\mathrm{M}}\boldsymbol{e}_i^{\mathrm{M}\prime})\end{pmatrix},
\end{aligned}
$$

then,

$$
\begin{aligned}
E_\xi[U_n(\boldsymbol{\beta})U_n'(\boldsymbol{\beta})] &= \mathrm{Var}_\xi[U_n(\boldsymbol{\beta})] = \mathrm{Var}_\xi\Big[\frac{1}{N}\sum_s w_i\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\Big] \\
&= \frac{1}{N^2}\sum_s w_i^2\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}\mathrm{Var}_\xi(\boldsymbol{y}_i-\boldsymbol{\mu}_i)V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{N^2}\sum_s w_i^2\frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}}V_i^{-1}\begin{pmatrix}E_\xi(\boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{o}\prime}) & E_\xi(\boldsymbol{e}_i^{\mathrm{o}}\boldsymbol{e}_i^{\mathrm{M}\prime})\\ E_\xi(\boldsymbol{e}_i^{\mathrm{M}}\boldsymbol{e}_i^{\mathrm{o}\prime}) & E_\xi(\boldsymbol{e}_i^{\mathrm{M}}\boldsymbol{e}_i^{\mathrm{M}\prime})\end{pmatrix}V_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}. \qquad (4.44)
\end{aligned}
$$

Once we subtract 4.44 from 4.43 we obtain the "meat" part in 4.18. This completes the proof. □

Estimation

It only remains to show that the bias of the naïve estimator of the "meat" of V_π (or equivalently, of V_{Sam}) is given by 4.22:

Proof of the bias 4.22. We have that:

$$\begin{aligned} z_k^* z_k^{*\prime} - z_k z_k' &= \frac{\partial \boldsymbol{\mu}_k'}{\partial B} V_k^{-1} \Big[\begin{pmatrix} \boldsymbol{e}_k^{\text{o}} \\ \boldsymbol{e}_k^{\text{I}} \end{pmatrix}^{\otimes 2} - \begin{pmatrix} \boldsymbol{e}_k^{\text{o}} \\ \boldsymbol{e}_k^{\text{M}} \end{pmatrix}^{\otimes 2} \Big] V_k^{-1} \frac{\partial \boldsymbol{\mu}_k}{\partial B} \\ &= \frac{\partial \boldsymbol{\mu}_k'}{\partial B} V_k^{-1} \Big[\begin{pmatrix} \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_k^{\text{o}\prime} & \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_k^{\text{I}\prime} \\ \boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_k^{\text{o}\prime} & \boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_k^{\text{I}\prime} \end{pmatrix} - \begin{pmatrix} \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_k^{\text{o}\prime} & \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_k^{\text{M}\prime} \\ \boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_k^{\text{o}\prime} & \boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_k^{\text{M}\prime} \end{pmatrix} \Big] V_k^{-1} \frac{\partial \boldsymbol{\mu}_k}{\partial B} \\ &= \frac{\partial \boldsymbol{\mu}_k'}{\partial B} V_k^{-1} \Big[\begin{pmatrix} \mathbf{0} & \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_k^{\text{I}\prime} \\ \boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_k^{\text{o}\prime} & \boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_k^{\text{I}\prime} \end{pmatrix} - \begin{pmatrix} \mathbf{0} & \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_k^{\text{M}\prime} \\ \boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_k^{\text{o}\prime} & \boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_k^{\text{M}\prime} \end{pmatrix} \Big] V_k^{-1} \frac{\partial \boldsymbol{\mu}_k}{\partial B}, \end{aligned}$$

so that ,

$$E_\xi [z_k^* z_k^{*\prime} - z_k z_k'] = \frac{\partial \boldsymbol{\mu}_k'}{\partial B} V_k^{-1} \Big[\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & E_\xi(\boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_k^{\text{I}\prime}) \end{pmatrix} - \begin{pmatrix} \mathbf{0} & E_\xi(\boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_k^{\text{M}\prime}) \\ E_\xi(\boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_k^{\text{o}\prime}) & E_\xi(\boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_k^{\text{M}\prime}) \end{pmatrix} \Big] V_k^{-1} \frac{\partial \boldsymbol{\mu}_k}{\partial B}. \quad (4.45)$$

Also, for $i \neq k$,

$$\begin{aligned} z_k^* z_i^{*\prime} - z_k z_i' &= \frac{\partial \boldsymbol{\mu}_k'}{\partial B} V_k^{-1} \Big[\begin{pmatrix} \boldsymbol{e}_k^{\text{o}} \\ \boldsymbol{e}_k^{\text{I}} \end{pmatrix} (\boldsymbol{e}_i^{\text{o}\prime},\, \boldsymbol{e}_i^{\text{I}\prime}) - \begin{pmatrix} \boldsymbol{e}_k^{\text{o}} \\ \boldsymbol{e}_k^{\text{M}} \end{pmatrix} (\boldsymbol{e}_i^{\text{o}\prime},\, \boldsymbol{e}_i^{\text{M}\prime}) \Big] V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial B} \\ &= \frac{\partial \boldsymbol{\mu}_k'}{\partial B} V_k^{-1} \Big[\begin{pmatrix} \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_i^{\text{o}\prime} & \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_i^{\text{I}\prime} \\ \boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_i^{\text{o}\prime} & \boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_i^{\text{I}\prime} \end{pmatrix} - \begin{pmatrix} \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_i^{\text{o}\prime} & \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_i^{\text{M}\prime} \\ \boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_i^{\text{o}\prime} & \boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_i^{\text{M}\prime} \end{pmatrix} \Big] V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial B} \\ &= \frac{\partial \boldsymbol{\mu}_k'}{\partial B} V_k^{-1} \Big[\begin{pmatrix} \mathbf{0} & \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_i^{\text{I}\prime} \\ \boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_i^{\text{o}\prime} & \boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_i^{\text{I}\prime} \end{pmatrix} - \begin{pmatrix} \mathbf{0} & \boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_i^{\text{M}\prime} \\ \boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_i^{\text{o}\prime} & \boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_i^{\text{M}\prime} \end{pmatrix} \Big] V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial B}, \end{aligned}$$

so that, for $i \neq k$,

$$E_\xi [z_k^* z_i^{*\prime} - z_k z_i'] = \frac{\partial \boldsymbol{\mu}_k'}{\partial B} V_k^{-1} \begin{pmatrix} \mathbf{0} & E_\xi(\boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_i^{\text{I}\prime}) \\ E_\xi(\boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_i^{\text{o}\prime}) & E_\xi(\boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_i^{\text{I}\prime}) \end{pmatrix} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial B}. \quad (4.46)$$

Therefore, the bias is given by $(n-1)^{-1}$ times the following quantity

$$\begin{aligned} &E_\xi \Big\{ \Big[(n-1) \sum_{k \in s} w_k^2 z_k^* z_k^{*\prime} - \sum_{k,i \in s} \sum_{i \neq k} w_k w_i z_k^* z_i^{*\prime} \Big] - \Big[(n-1) \sum_{k \in s} w_k^2 z_k z_k' - \sum_{k,i \in s} \sum_{i \neq k} w_k w_i z_k z_i' \Big] \Big\} \\ &= E_\xi \Big\{ (n-1) \sum_{k \in s} w_k^2 (z_k^* z_k^{*\prime} - z_k z_k') - \sum_{k,i \in s} \sum_{i \neq k} w_k w_i (z_k^* z_i^{*\prime} - z_k z_i') \Big\} \\ &= (n-1) \sum_{k \in s} w_k^2 \frac{\partial \boldsymbol{\mu}_k'}{\partial B} V_k^{-1} \Big[\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & E_\xi(\boldsymbol{e}_k^{\text{I}} \boldsymbol{e}_k^{\text{I}\prime}) \end{pmatrix} - \begin{pmatrix} \mathbf{0} & E_\xi(\boldsymbol{e}_k^{\text{o}} \boldsymbol{e}_k^{\text{M}\prime}) \\ E_\xi(\boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_k^{\text{o}\prime}) & E_\xi(\boldsymbol{e}_k^{\text{M}} \boldsymbol{e}_k^{\text{M}\prime}) \end{pmatrix} \Big] V_k^{-1} \frac{\partial \boldsymbol{\mu}_k}{\partial B} \end{aligned}$$

$$-\sum_{k,i\in s}\sum_{i\neq k} w_k w_i \frac{\partial \boldsymbol{\mu}_k'}{\partial B} V_k^{-1} \begin{pmatrix} \mathbf{0} & E_\xi(\boldsymbol{e}_k^{\mathrm{o}} \boldsymbol{e}_i^{\mathrm{I}\prime}) \\ E_\xi(\boldsymbol{e}_k^{\mathrm{I}} \boldsymbol{e}_i^{\mathrm{o}\prime}) & E_\xi(\boldsymbol{e}_k^{\mathrm{I}} \boldsymbol{e}_i^{\mathrm{I}\prime}) \end{pmatrix} V_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial B}.$$

□

Chapter 5

Simulation Studies

5.1 Setup for Continuous Response

The Model Used for Simulations

In this chapter we use the first four cycles of the NLSCY dataset, together with its synthetic files, to build up the models and the simulations. The response variable, or variable of interest, is the physical aggression score (PAS) of the kid, as defined in section 1.2, and, in this section, treated as a continuous variable. These data have been analyzed by Carrillo et al. (2005); Carrillo-Garcia (2006); Carrillo et al. (2006). They found that, among the explanatory variables that they considered, the only ones that turned out to be significant were the age of the kid (AGE), the squared of the age (AGE^2), the depression score of the person most knowledgeable about the kid (DeprePMK), the punitive/aversive parenting score (Punitive), and the child's gender (GENDER).

In order to reduce the size of our simulation experiments, we decided to drop one of the least significant variables; this could be either DeprePMK or Punitive. DeprePMK is a score from 0 to 36 and a high score indicates the presence of depression symptoms; whereas Punitive is a score ranging from 0 to 19 with a high score indicating punitive/aversive interactions between parent and kid. We would like to have either of these variables categorized into three categories, and so, have $4\times3\times2 = 24$ (AGE$\times$Punitive/DeprePMK$\times$GENDER) "imputation classes" in total. We found that the variable DeprePMK is easier than Punitive to categorize into three classes and have the three classes have about the same number of respondents. Then we decided to drop the variable Punitive and keep the variable DeprePMK.

We use the following four explanatory variables. AGE, which at cycle 1 has four categories (2, 3, 4, and 5); AGE^2; DeprePMK, with three categories (0, 3, and 9); and GENDER, with two categories (0 and 1). In this chapter we treat the four explanatory variables as continuous for the main model, but as categorical for the imputation procedure; and the response variable, PAS, as a continuous response (in the present section). We have four categories of AGE at each cycle, three categories of DeprePMK, and two categories of GENDER; this gives as $4\times3\times2 = 24$ imputation classes at each cycle.

The true model parameters used in the simulations are obtained as follows. We consider

the linear model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3} + \varepsilon_{ij}, \tag{5.1}$$

where

$$\begin{aligned} Y_{ij} &: \text{ PAS of the } i\text{-th subject at } j\text{-th time,} \\ x_{ij1} &= \text{Age of subject } i \text{ at time } j, \\ x_{ij2} &= \text{Depression score of the PMK of subject } i \text{ at time } j, \\ x_{i3} &= \text{Gender of } i\text{-th subject,} \\ \varepsilon_i &= (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4}) \overset{\text{ind.}}{\backsim} (\mathbf{0}, \sigma^2 \mathbf{R}). \end{aligned}$$

The Pseudo-GEE method is then applied to the NLSCY dataset, using model 5.1, with the correlation structure unspecified and estimated using the method of moments, parametrized by the Pearson correlation, as in Liang and Zeger (1986). The estimates for the regression coefficients are:

$$\beta_0 = 5.6225,\ \beta_1 = -1.0982,\ \beta_2 = 0.0656,\ \beta_3 = 0.0609,\ \beta_4 = -0.2900;$$

with auto-correlation matrix:

$$\mathbf{R} = \begin{pmatrix} 1 & 0.4123 & 0.3919 & 0.3353 \\ 0.4123 & 1 & 0.4798 & 0.3172 \\ 0.3919 & 0.4798 & 1 & 0.4370 \\ 0.3353 & 0.3172 & 0.4370 & 1 \end{pmatrix},$$

and dispersion parameter:

$$\phi = \sigma^2 = 3.66842.$$

We will generate data from model 5.1 to examine the statistical properties of the procedures we have developed. This model was generated from the real NLSCY dataset, and so, can be thought of as a reasonable model. Nonetheless, we will not use the real NLSCY dataset for our simulations because, due to privacy concerns, it does not give us the freedom to manipulate the data at our will on our computing resources. For the simulations we use a subset of the synthetic NLSCY data, which was released by Statistics Canada for the 2005 SSC case study (together with model 5.1).

In this dataset there are 458 kids with the three variables, AGE, DeprePMK, and GENDER, all completed at all four cycles. We replicate these 458 subjects 40 times, to generate an artificial finite population of 18,320 subjects. Notice that so far we have only generated the covariates, but not the response variable. The covariates will remain fixed throughout all the simulations, whereas the response variable will be re-generated on each simulation. To generate the response variable we first make use of the true model 5.1 to get the mean response for each subject at each cycle. This model mean also stays fixed for the whole process since it is a function merely of the covariates. The following equations summarize the mean generation for the 18,320 subjects:

$$\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \mu_{i4}), \tag{5.2}$$

and $\mu_{ij} = E_\xi(Y_{ij}|\mathbf{x}_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3}$, $i = 1, 2, \cdots, N$, $j = 1, 2, 3, 4$; where $\mathbf{x}_{ij} = (x_{ij1}, x_{ij1}^2, x_{ij2}, x_{i3})$.

The Imputation Classes

We create four variables (one for each cycle) defining the 24 "imputation cells" for each of the four cycles. These cells are defined by the cross-classification of the 4 categories of AGE, 3 categories of DeprePMK, and 2 categories of GENDER. Since DeprePMK is a time varying covariate, the imputation cells change from cycle to cycle.

The Sampling Schemes

We now select the samples and generate the response variable [1]. We select 1,000 samples of each of the sizes $n = 120, 240, 360, 480, 600, 720, 840, 960, 1080$, and 1200; this gives us selected samples of about $5, 10, 15, 20, 25, 30, 35, 40, 45$, or 50 in each imputation class.

For SRS we first create a "weight" variable for each of the $N = 18,320$ subjects, which is equal to $18,320/n$. Then we select without replacement a sample of n numbers from 1 to 18,320, and take as sample the n subjects corresponding to the selected numbers. For each of the n subjects in the sample we generate the four (auto-correlated) responses in the following way:

$$(y_{i1}, y_{i2}, y_{i3}, y_{i4})' = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \mu_{i4})' + [\mathrm{MVN}\big((0,0,0,0), \mathrm{I}_4\big) \times \big(\texttt{mroot}(\phi\mathbf{R})\big)'].$$

For stratified random sampling (STSI), on the other hand, we create a stratum indicator variable which is equal to 1 if AGE1 (the age at cycle 1) is 2 or 3, and equal to 2 if AGE1 is 4 or 5; we obtain $N_1 = 9,000$ and $N_2 = 9,320$. This is, we stratify based on age. We allocate $n_1 = n/3$ units to stratum 1, and $n_2 = 2n/3$ to stratum 2. Our stratified samples will have twice as many units in stratum 2 as those in stratum 1. This allocation aims to be "y-proportional;" the coefficient of variation of the PAS has a tendency to increase with age. Then we create a "weight" variable for each of the 18,320 subjects, which is equal to $9000/n_1$ for subjects in stratum 1, and $9320/n_2$ for those in stratum 2. After that, we select without replacement a sample of n_1 elements from stratum 1, and a sample of n_2 elements from stratum 2. Finally, for each of the $n = n_1 + n_2$ subjects in the sample we generate the four (auto-correlated) responses exactly as for the SRS case.

For cluster sampling with clusters selected by simple random sampling (SIC), we first create artificial clusters, of sizes 5 or 10, in the finite population as follows. We group randomly the 18,320 subjects into 2,748 clusters; 1,832 of size 5 and 916 of size 10. Note that here the clusters do not change over time. For each cluster c and cycle j, we generate a random effect,

[1] In theory we have to generate the response variable for each of the 18,320 subjects before each and every time we select a new sample. But in practice we only need the response variable for the subjects actually selected in a given sample.

$b_{cj} \sim N(0,1)$ independently. Then, for any subject i in cluster c and cycle j, we redefine $\mu_{ijc} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3} + b_{cj}$, were the β's and $\mathbf{x}_{ij}$ are as before. Since $\phi = \sigma^2 = 3.66842$, this produces a correlation of 0.2142 between any two subjects in the same cluster and cycle. Now, if our target sample of elements is of size n, then we select a sample s_I of $n_I = n/6.6666666$ clusters at random, and include in the sample all the elements belonging to the selected clusters. The element sample size[2] is then random, but on average it is n. Also, on average the number of clusters of size 5 is $2n_I/3$ and of size 10 is $n_I/3$. Finally, for each subject in the sample we generate the four responses exactly as for the SRS and STSI cases, but with the μ_{ijc} as defined here.

The Respondent/Non-respondent Sets

Next we "create" the respondents and non-respondents for each cycle, among the selected individuals. We use the following probabilities of non-response: $p_m = 0.05, 0.10, 0.15, 0.20,$ and 0.25. This will produce anywhere from $r = 90$ to 1140 actual respondents at each cycle for the full sample, or anywhere from 3.75 to 47.5 actual respondents in each imputation cell, on average.

For the MAR mechanism, in each of the 24 cells and for each of the 4 cycles we choose, at random, a missing probability (p_m) according to the desired overall missing probability, as follows:

- from {.03, .04, .05, .06, .07} if the overall missing probability is 0.05
- from {.08, .09, .10, .11, .12} if the overall missing probability is 0.10
- from {.13, .14, .15, .16, .17} if the overall missing probability is 0.15
- from {.18, .19, .20, .21, .22} if the overall missing probability is 0.20, and
- from {.23, .24, .25, .26, .27} if the overall missing probability is 0.25.

Then, in each cell-time we create a vector "one.zeros" of $n_{\text{cell}} \times p_m$ zeros and $n_{\text{cell}} \times (1 - p_m)$ ones. And finally we generate a variable which represents the missing status and corresponds to the vector one.zeros sorted randomly within cell-times.

We generate the respondents/nonrespondents in this way because we do not want to be too restrictive about the missing fraction. For example, when we want the overall missing percentage to be 15%, we do not want it to be *exactly 15% in every single cell*; on the contrary, we would like to allow this percentage to vary somewhat from cell to cell. So that the setup resembles more closely what happens in practice.

Imputation for Missing Values

The last step of the setup is to select the donors for the nonrespondents. For the unweighted hot-deck, in each of the 24 imputation cells (independently), and for each of the 4 cycles

[2] Number of actual subjects selected.

(independently), we select *with replacement* a number of respondents equal to the number of non-respondents, and fill in the missing values of PAS with those of the selected respondents.

Estimation

With regard to the estimation procedure, since we are now dealing with a continuous response, the point estimator in section 4.2 is obtained as follows. We let as starting values:

$$\boldsymbol{\beta}^{(0)} = \Big(\sum_s w_i X_i X_i'\Big)^{-1} \sum_s w_i X_i \boldsymbol{y}_i^*;$$

letting $e_{it}^{(0)} = y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}^{(0)}$ and

$$R_{it} = \begin{cases} 1 \text{ if subject } i \text{ observed at cycle } t \\ 0 \text{ if subject } i \text{ missing at cycle } t, \end{cases}$$

then

$$\phi^{(0)} = \sigma^{2(0)} = \frac{\sum_{i\in s}\sum_{t=1}^4 w_i R_{it} e_{it}^{2(0)}}{\sum_{i\in s}\sum_{t=1}^4 w_i R_{it} - p}, \quad \hat{\alpha}_{tt'} = \widehat{\text{corr}}(y_{it}, y_{it'}) = \frac{\sum_{i\in s} w_i R_{it} R_{it'} e_{it}^{(0)} e_{it'}^{(0)}}{\phi^{(0)}[\sum_{i\in s} w_i R_{it} R_{it'} - p]}, \quad (5.3)$$

and

$$\mathbf{R}^{(0)} = \begin{pmatrix} 1 & \hat{\alpha}_{12} & \hat{\alpha}_{13} & \hat{\alpha}_{14} \\ \hat{\alpha}_{12} & 1 & \hat{\alpha}_{23} & \hat{\alpha}_{24} \\ \hat{\alpha}_{13} & \hat{\alpha}_{23} & 1 & \hat{\alpha}_{34} \\ \hat{\alpha}_{14} & \hat{\alpha}_{24} & \hat{\alpha}_{34} & 1 \end{pmatrix}.$$

And then we reiterate; while the maximum difference between $\boldsymbol{\beta}^{(l-1)}$ and $\boldsymbol{\beta}^{(l)}$ is bigger than 0.00005 we do:

$$\boldsymbol{\beta}^{(l+1)} = \Big(\sum_s w_i X_i [\mathbf{R}^{(l)}]^{-1} X_i'\Big)^{-1} \sum_s w_i X_i [\mathbf{R}^{(l)}]^{-1} \boldsymbol{y}_i^*,$$

and $\phi^{(l+1)}$ and $\mathbf{R}^{(l+1)}$ take the same form.

After convergence we obtain $\hat{\boldsymbol{\beta}}$, $\hat{\phi}$ and $\hat{\mathbf{R}}$. Then we estimate the total variance using the methodology developed in section 4.3. For the linear model 5.1 and for the sampling schemes considered here, some of the involved terms can be spelled out as follows.

The "bread" term of all the variance components takes the form $N\hat{\phi}\left[\sum_s w_i X_i \hat{\mathbf{R}}^{-1} X_i'\right]^{-1}$. We estimate the inner piece of V_{Imp} with:

$$meat(\hat{V}_{\text{Imp}}) = \frac{1}{N^2\hat{\phi}^2}\Big\{\sum_s w_i^2 X_i \hat{\mathbf{R}}^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \text{diag}(s_{\tau r}^2) \end{pmatrix} \hat{\mathbf{R}}^{-1} X_i' + \Big(\sum_s w_i X_i \hat{\mathbf{R}}^{-1} [\begin{pmatrix} \boldsymbol{y}_i^o \\ \bar{\mathbf{y}}_{\tau r} \end{pmatrix} - X_i'\hat{\boldsymbol{\beta}}]\Big)^{\otimes 2}\Big\}.$$

Let $\boldsymbol{r}_k^* = \boldsymbol{y}_k^* - X_k'\hat{\boldsymbol{\beta}}$. For SRS, the naïve estimator of the inner piece of V_{Sam} is:

$$meat(\hat{V}_\pi^*) = \frac{1}{N^2\hat{\phi}^2(n-1)}\Big\{n\sum_{k\in s}[w_k X_k \hat{\mathbf{R}}^{-1}\boldsymbol{r}_k^*]^{\otimes 2} - \Big(\sum_{i\in s} w_i X_i \hat{\mathbf{R}}^{-1}\boldsymbol{r}_i^*\Big)^{\otimes 2}\Big\};$$

for STSI, it is:

$$meat(\hat{V}_\pi^*) = \frac{1}{N^2\hat{\phi}^2}\sum_{h=1}^{2}\frac{1}{n_h-1}\Big\{n_h\sum_{k\in s_h}[w_k X_k\hat{\mathbf{R}}^{-1}\boldsymbol{r}_k^*]^{\otimes 2} - \Big(\sum_{i\in s_h} w_i X_i\hat{\mathbf{R}}^{-1}\boldsymbol{r}_i^*\Big)^{\otimes 2}\Big\};$$

and for SIC:

$$meat(\hat{V}_\pi^*) = \frac{1}{N^2\hat{\phi}^2(n_I-1)}\Big\{n_I\sum_{k\in s_I}\Big[\sum_{l\in U_k} w_l X_l\hat{\mathbf{R}}^{-1}\boldsymbol{r}_l^*\Big]^{\otimes 2} - \Big(\sum_{k\in s_I}\sum_{l\in U_k} w_l X_l\hat{\mathbf{R}}^{-1}\boldsymbol{r}_l^*\Big)^{\otimes 2}\Big\}.$$

Now let $\boldsymbol{r}_i^{\mathrm{I}} = \boldsymbol{y}_i^{\mathrm{I}} - X_i^{\mathrm{M}\prime}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{r}_i^{\mathrm{o}} = \boldsymbol{y}_i^{\mathrm{o}} - X_i^{\mathrm{o}\prime}\hat{\boldsymbol{\beta}}$. The inner piece of the estimated bias, in eq. 4.23, for SRS is:

$$meat(Bias[\hat{V}_\pi^*]) = \frac{1}{N^2\hat{\phi}^2}\Bigg\{\underbrace{\sum_{k\in s} w_k^2 X_k\hat{\mathbf{R}}^{-1}\Big[\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_k^{\mathrm{I}\prime}\end{pmatrix} - \hat{\phi}\begin{pmatrix}\mathbf{0} & \hat{\mathbf{R}}^{(\mathrm{oM})}\\ \hat{\mathbf{R}}^{(\mathrm{oM})\prime} & \hat{\mathbf{R}}^{(\mathrm{MM})}\end{pmatrix}\Big]\hat{\mathbf{R}}^{-1}X_k'}_{\text{single summation term}}$$
$$\underbrace{-\frac{1}{n-1}\sum_{k,i\in s}\sum_{i\neq k} w_k w_i X_k\hat{\mathbf{R}}^{-1}\begin{pmatrix}\mathbf{0} & \boldsymbol{r}_k^{\mathrm{o}}\boldsymbol{r}_i^{\mathrm{I}\prime}\\ \boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_i^{\mathrm{o}\prime} & \boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_i^{\mathrm{I}\prime}\end{pmatrix}\hat{\mathbf{R}}^{-1}X_i'}_{\text{double summation term}}\Bigg\};$$

for STSI:

$$meat(Bias[\hat{V}_\pi^*]) = \frac{1}{N^2\hat{\phi}^2}\sum_{h=1}^{2}\Bigg\{\underbrace{\sum_{k\in s_h} w_k^2 X_k\hat{\mathbf{R}}^{-1}\Big[\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_k^{\mathrm{I}\prime}\end{pmatrix} - \hat{\phi}\begin{pmatrix}\mathbf{0} & \hat{\mathbf{R}}^{(\mathrm{oM})}\\ \hat{\mathbf{R}}^{(\mathrm{oM})\prime} & \hat{\mathbf{R}}^{(\mathrm{MM})}\end{pmatrix}\Big]\hat{\mathbf{R}}^{-1}X_k'}_{\text{single summation term}}$$
$$\underbrace{-\frac{1}{n_h-1}\sum_{k,i\in s_h}\sum_{i\neq k} w_k w_i X_k\hat{\mathbf{R}}^{-1}\begin{pmatrix}\mathbf{0} & \boldsymbol{r}_k^{\mathrm{o}}\boldsymbol{r}_i^{\mathrm{I}\prime}\\ \boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_i^{\mathrm{o}\prime} & \boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_i^{\mathrm{I}\prime}\end{pmatrix}\hat{\mathbf{R}}^{-1}X_i'}_{\text{double summation term}}\Bigg\};$$

and for SIC:

$$\begin{aligned} meat(Bias[\hat{V}_\pi^*]) = \\ \frac{1}{N^2\hat{\phi}^2}\Bigg[&\underbrace{\sum_{k\in s_I}\sum_{l\in U_k} w_l^2 X_l\hat{\mathbf{R}}^{-1}\Big[\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{r}_l^{\mathrm{I}}\boldsymbol{r}_l^{\mathrm{I}\prime}\end{pmatrix} - \hat{\phi}\begin{pmatrix}\mathbf{0} & \hat{\mathbf{R}}^{(\mathrm{oM})}\\ \hat{\mathbf{R}}^{(\mathrm{oM})\prime} & \hat{\mathbf{R}}^{(\mathrm{MM})}\end{pmatrix}\Big]\hat{\mathbf{R}}^{-1}X_l'}_{\text{single summation term}} \\ &\underbrace{+\sum_{k\in s_I}\sum_{l,m\in U_k}\sum_{m\neq l} w_l w_m X_l\hat{\mathbf{R}}^{-1}\begin{pmatrix}\mathbf{0} & \boldsymbol{r}_l^{\mathrm{o}}\boldsymbol{r}_m^{\mathrm{I}\prime}\\ \boldsymbol{r}_l^{\mathrm{I}}\boldsymbol{r}_m^{\mathrm{o}\prime} & \boldsymbol{r}_l^{\mathrm{I}}\boldsymbol{r}_m^{\mathrm{I}\prime}\end{pmatrix}\hat{\mathbf{R}}^{-1}X_m'}_{\text{double summation}} \\ &\underbrace{-\frac{1}{n_I-1}\sum_{k,i\in s_I}\sum_{i\neq k}\sum_{l\in U_k}\sum_{m\in U_i} w_l w_m X_l\hat{\mathbf{R}}^{-1}\begin{pmatrix}\mathbf{0} & \boldsymbol{r}_l^{\mathrm{o}}\boldsymbol{r}_m^{\mathrm{I}\prime}\\ \boldsymbol{r}_l^{\mathrm{I}}\boldsymbol{r}_m^{\mathrm{o}\prime} & \boldsymbol{r}_l^{\mathrm{I}}\boldsymbol{r}_m^{\mathrm{I}\prime}\end{pmatrix}\hat{\mathbf{R}}^{-1}X_m'}_{\text{between cluster (double summation)}}\Bigg]. \end{aligned} \quad \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\}\text{within cluster} \tag{5.4}$$

Finally, for SRS, the estimator of the middle part of $C_{\text{Imp-Sam}}$ reduces to:

$$meat(\hat{C}_{\text{Imp-Sam}}) = \frac{1}{N^2\hat{\phi}^2}\Bigg\{\underbrace{\sum_{i,j\in s}\sum_{i\neq j} w_i w_j X_i \hat{\mathbf{R}}^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \boldsymbol{r}_i^{\text{I}}\boldsymbol{r}_j^{\text{o}\prime} & \mathbf{0}\end{pmatrix}\hat{\mathbf{R}}^{-1}X_j'}_{\text{double summation term}}$$
$$\underbrace{-\sum_{s} w_i^2 X_i \hat{\mathbf{R}}^{-1}\hat{\phi}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \hat{\mathbf{R}}^{(\text{oM})\prime} & \hat{\mathbf{R}}^{(\text{MM})}\end{pmatrix}\hat{\mathbf{R}}^{-1}X_i'}_{\text{single summation term}}\Bigg\};$$

for STSI it takes the following form:

$$meat(\hat{C}_{\text{Imp-Sam}}) = \frac{1}{N^2\hat{\phi}^2}\sum_{h=1}^{2}\Bigg\{\underbrace{\sum_{i,j\in s_h}\sum_{i\neq j} w_i w_j X_i \hat{\mathbf{R}}^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \boldsymbol{r}_i^{\text{I}}\boldsymbol{r}_j^{\text{o}\prime} & \mathbf{0}\end{pmatrix}\hat{\mathbf{R}}^{-1}X_j'}_{\text{double summation term}}$$
$$\underbrace{-\sum_{s_h} w_i^2 X_i \hat{\mathbf{R}}^{-1}\hat{\phi}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \hat{\mathbf{R}}^{(\text{oM})\prime} & \hat{\mathbf{R}}^{(\text{MM})}\end{pmatrix}\hat{\mathbf{R}}^{-1}X_i'}_{\text{single summation term}}\Bigg\};$$

and for SIC:

$$meat(\hat{C}_{\text{Imp-Sam}}) = \frac{1}{N^2\hat{\phi}^2}\Bigg\{\underbrace{\sum_{k\in s_I}\sum_{m\in s_I}\sum_{i\in U_k}\sum_{\substack{j\in U_m\\ j\neq i}} w_i w_j X_i \hat{\mathbf{R}}^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \boldsymbol{r}_i^{\text{I}}\boldsymbol{r}_j^{\text{o}\prime} & \mathbf{0}\end{pmatrix}\hat{\mathbf{R}}^{-1}X_j'}_{\text{double summation term}}$$
$$\underbrace{-\sum_{k\in s_I}\sum_{i\in U_k} w_i^2 X_i \hat{\mathbf{R}}^{-1}\hat{\phi}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \hat{\mathbf{R}}^{(\text{oM})\prime} & \hat{\mathbf{R}}^{(\text{MM})}\end{pmatrix}\hat{\mathbf{R}}^{-1}X_i'}_{\text{single summation term}}\Bigg\}.$$

The alternative variance estimation of section 4.4 takes the following form. The inner part of the imputation variance component is estimated by:

$$\widehat{\text{Var}_I}[U_n^*(\boldsymbol{\beta})] = \frac{1}{N^2\hat{\phi}^2}\sum_{s} w_i^2 X_i \hat{\mathbf{R}}^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \text{diag}(s_{\tau r}^2)\end{pmatrix}\hat{\mathbf{R}}^{-1}X_i'.$$

The inner part of the sampling variance component is estimated by $(n/r)^2\hat{V}_{\tau\text{naïve}}$, where, for SRS,

$$\hat{V}_{\tau\text{naïve}} = \frac{1}{N^2\hat{\phi}^2(n-1)}\Big[n\sum_{k\in s} w_k^2\hat{\boldsymbol{z}}_{\tau k}\hat{\boldsymbol{z}}_{\tau k}' - \Big(\sum_{i\in s} w_i\hat{\boldsymbol{z}}_{\tau i}\Big)^{\otimes 2}\Big];$$

for STSI:

$$\hat{V}_{\tau\text{naïve}} = \frac{1}{N^2\hat{\phi}^2}\sum_{h=1}^{2}\frac{1}{n_h-1}\Big[n_h\sum_{k\in s_h} w_k^2\hat{\boldsymbol{z}}_{\tau k}\hat{\boldsymbol{z}}_{\tau k}' - \Big(\sum_{i\in s_h} w_i\hat{\boldsymbol{z}}_{\tau i}\Big)^{\otimes 2}\Big];$$

and for SIC:

$$\hat{V}_{\tau\text{naïve}} = \frac{1}{N^2\hat{\phi}^2(n_I-1)}\left[n_I \sum_{k\in s_I}\Big[\sum_{l\in U_k} w_l\hat{z}_{\tau l}\Big]^{\otimes 2} - \Big(\sum_{i\in s_I}\sum_{l\in U_i} w_l\hat{z}_{\tau l}\Big)^{\otimes 2}\right];$$

and

$$\hat{z}_{\tau i} = X_i\hat{\mathbf{R}}^{-1}\begin{pmatrix} y_i^{\mathrm{o}} - X_i^{\mathrm{o}\prime}\hat{\boldsymbol{\beta}} \\ \bar{y}_{\tau r} - X_i^{\mathrm{M}\prime}\hat{\boldsymbol{\beta}}\end{pmatrix}.$$

5.2 Setup for Binary Response

The Simulation Model

We are interested in evaluating our method of chapter 4 for binary responses; because in surveys there are usually many binary outcomes of interest. Here we use the NLSCY dataset to build our true logistic regression model, and the synthetic files for the simulations. In this part of the thesis the response variable, Y, is the child's physical aggression score (PAS), but recoded into two categories of roughly equal number of subjects. A category of those with "low" PAS, for which the original PAS is 1.5 or less, and recoded as "0"; and a category of those with "high" PAS, for which the original PAS was bigger than 1.5, recoded as "1".

We include as explanatory variables those found to be significant by Carrillo et al. (2005), Carrillo-Garcia (2006), and Carrillo et al. (2006): the child's age (AGE), with four categories at cycle 1 (2, 3, 4, and 5); the age squared (AGE^2), the depression score of the person most knowledgeable about the kid (DeprePMK), with three categories (0, 3, and 9); and the child's gender (GENDER), with two categories (0 and 1). Even though the punitive/aversive parenting score was found to be significant in their studies, we decided not to include it here to keep the size of our simulations manageable; this variable was one of the least significant ones. These covariates are considered to be continuous for the main model, but categorical for the hot-deck imputation. There are $4\times 3\times 2 = 24$ imputation classes at each cycle (4 categories of AGE at each cycle $\times$ 3 categories of DeprePMK $\times$ 2 categories of GENDER).

To get our "true" model, we fit a logistic Pseudo-GEE model to the 3,049 subjects in the complete NLSCY dataset, using the funnel weights. We use an unspecified correlation structure, but estimate it with the odds ratio parametrization. Song (2007) points out that "to measure dependence between nonnormal variables, there are some better tools than Pearson correlation. For example, odds ratio (OR) is a measure of association for categorical variates." Lipsitz et al. (1991), Liang et al. (1992), and Carey et al. (1993) have used odds ratios to measure the association among binary and other categorical data. They, and other references in those manuscripts, argue that for binary responses, the odds ratio has some desirable properties and is easier to interpret than the correlation coefficient.

We consider the following superpopulation logistic regression model, which we use to generate our finite population with a binary response,

$$\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3}, \tag{5.5}$$

where

$$\begin{aligned}
p_{ij} &= P(Y_{ij}=1|\mathbf{x}_{ij}),\\
Y_{ij} &: \text{ PAS of the } i\text{-th subject at } j\text{-th time},\\
\mathbf{x}_{ij} &= (x_{ij1}, x_{ij1}^2, x_{ij2}, x_{i3}),\\
x_{ij1} &= \text{Age of subject } i \text{ at time } j,\\
x_{ij2} &= \text{Depression score of the PMK of subject } i \text{ at time } j,\\
x_{i3} &= \text{Gender of } i\text{-th subject.}
\end{aligned}$$

The true values of the model parameters β_0, β_1, β_2, β_3, β_4 (and also the odds ratios) are set to be the estimated regression coefficients from fitting the logistic regression model 5.5 using the completed NLSCY dataset, and are given by

$$\beta_0 = 2.7181,\ \beta_1 = -0.8959,\ \beta_2 = 0.0530,\ \beta_3 = 0.0701,\ \beta_4 = -0.2811;$$

and odds ratios among the four responses:

$$\begin{aligned}
&OR_{12} = 4.7669,\ OR_{13} = 3.9257,\ OR_{14} = 3.0930,\\
&OR_{23} = 5.8401,\ OR_{24} = 4.4069,\ OR_{34} = 6.6430,
\end{aligned} \tag{5.6}$$

where OR_{st} is the odds ratio between responses at times s and t; the dispersion parameter for this case is $\phi = 1$.

We treat the model in 5.5 as the true model; all our simulation results are evaluated in comparison to it. But for the simulations we use the synthetic NLSCY dataset since we do not have permission to manipulate the real NLSCY dataset on our computing resources. The artificial finite population, consisting of 18,320 subjects, is created by repeating 40 times the 458 kids with the three covariates, AGE, DeprePMK, and GENDER, fully observed. The covariates remain fixed for all the simulations, whereas we re-generate the response variable on each simulation.

With the help of the true model 5.5 we get the mean response for each subject at each cycle. This model mean does not change from simulation to simulation since it is a function of the covariates. The mean response for each subject is generated as follows:

$$\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \mu_{i3}, \mu_{i4}), \tag{5.7}$$

and $\mu_{ij} = \{1 + \exp(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3})\}^{-1}$, $i = 1, 2, \cdots, N$, $j = 1, 2, 3, 4$.

Imputation and Sampling Schemes

The 24 "imputation cells" for each cycle are created by the cross-classification of the 4 categories of AGE, 3 categories of DeprePMK, and 2 categories of GENDER. These imputation classes are time varying because DeprePMK is time varying.

We now select 1,000 samples of each of the sizes $n = 120$, 240, 360, 480, 600, 720, 840, 960, 1080, and 1200; this gives us selected samples of about $5, 10, 15, 20, 25, 30, 35, 40, 45$, or 50 in each imputation class.

For simple random sampling (SRS), we first create a "weight" variable for each of the $N = 18,320$ subjects, which is equal to $18,320/n$. Then we select without replacement a sample of n subjects from among the 18,320 in the finite population. For stratified random sampling (STSI), on the other hand, we create a stratum indicator variable, equal to 1 if AGE1 is 2 or 3, and equal to 2 if AGE1 is 4 or 5 ($N_1 = 9,000$, $N_2 = 9,320$); this is, we stratify based on age at wave 1. We allocate $n_1 = n/3$ units to stratum 1, and $n_2 = 2n/3$ to stratum 2. So, we select twice as many units in stratum 2 as in stratum 1. We chose to allocate in this way because the coefficient of variation of the PAS increases with age. In this case the weight is equal to $9000/n_1$ for stratum 1, and $9320/n_2$ for stratum 2. After that, we select without replacement a sample of n_1 elements from stratum 1, and a sample of n_2 elements from stratum 2.

In each selected sample we next have to generate the four (auto-correlated) binary responses $(y_1, y_2, y_3, y_4)'$ for each of the n subjects. This responses must have the means 5.7, and satisfy the odds ratios 5.6. We use the method based on Gaussian copula, presented by Song (2000), to generate these responses. Another possible method to generate correlated responses from some exponential family distributions is the one discussed in Song (1997); however, this method is not well suited for binary variables.

The setup for the SIC case is more complicated precisely because we need to generate the "clustering" in the population somehow. We first group randomly the 18,320 subjects in the population into 2,748 clusters, 916 of size 10 and 1,832 of size 5. Then we add a random effect $b_{cj} \overset{\text{iid}}{\backsim} N(0, 0.2)$ to the linear predictor $\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3}$ for every subject i in cluster c at time j. The clusters do not change from cycle to cycle but the random effect does. Thus for a subject i in cluster c at time j we have $\mu_{ijc} = \{1 + \exp(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2 + \beta_3 x_{ij2} + \beta_4 x_{i3} + b_{cj})\}^{-1}$ and $\boldsymbol{\mu}_i$ as in 5.7. Then we generate the four auto-correlated binary responses $(y_1, y_2, y_3, y_4)'$ for each of the N subjects in the population satisfying the odds ratios in 5.6 by the method in Song (2000). We keep the same responses for all N subjects fixed throughout the 1,000 simulations.

Finally, for any of the three sampling designs, the "creation" of respondents and non-respondents, and the selection of donors is carried out exactly as in the case of continuous response, detailed in section 5.1.

Estimation

With regard to the estimation procedure in each selected sample, since we are now dealing with logistic regression for a binary response, and odds ratio parametrization for correlation, we proceed as follows. The point estimator in section 4.2 is obtained as follows. Our initial value is $\boldsymbol{\beta}^{(0)} = \left(\beta_0^{(0)}, 0, 0, 0, 0\right)'$, where

$$\beta_0^{(0)} = \log\left(\frac{\sum_{i\in s}\sum_{t=1}^{4} w_i R_{it} y_{it}}{\sum_{i\in s}\sum_{t=1}^{4} w_i R_{it}(1-y_{it})}\right),$$

where R_{it} is the response indicator for subject i at time t. In other words, $\beta_0^{(0)}$ is the estimate of the log odds of high PAS, collapsing all four cycles of responses and ignoring all covariates.

We also get estimates for the six odds ratios:

$$\widehat{OR}_{st} = \frac{\sum_{i\in s} w_i R_{it} R_{is} y_{it} y_{is} \cdot \sum_{i\in s} w_i R_{it} R_{is} (1-y_{it})(1-y_{is})}{\sum_{i\in s} w_i R_{it} R_{is} y_{it} (1-y_{is}) \cdot \sum_{i\in s} w_i R_{it} R_{is} (1-y_{it}) y_{is}}, \tag{5.8}$$

for $st = 12, 13, 14, 23, 24$, and 34. Here $\widehat{OR}_{st}$ is an estimate of the odds ratio between responses at times s and t.

Then, while the maximum difference between $\boldsymbol{\beta}^{(l-1)}$ and $\boldsymbol{\beta}^{(l)}$ is bigger than 0.00005 we reiterate the following steps. First we get an estimate of the matrix $\mathbf{R}_i$ for each subject, based on the current estimate $\boldsymbol{\beta}^{(l)}$, as:

$$\mathbf{R}_i^{(l)} = \begin{pmatrix} 1 & \hat{\alpha}_{i12} & \hat{\alpha}_{i13} & \hat{\alpha}_{i14} \\ \hat{\alpha}_{i12} & 1 & \hat{\alpha}_{i23} & \hat{\alpha}_{i24} \\ \hat{\alpha}_{i13} & \hat{\alpha}_{i23} & 1 & \hat{\alpha}_{i34} \\ \hat{\alpha}_{i14} & \hat{\alpha}_{i24} & \hat{\alpha}_{i34} & 1 \end{pmatrix}; \tag{5.9}$$

where

$$\hat{\alpha}_{ist} = \widehat{\text{corr}}(Y_{is}, Y_{it}) = \frac{\hat{p}_{ist} - \hat{\mu}_{is}\hat{\mu}_{it}}{\sqrt{\hat{\mu}_{is}(1-\hat{\mu}_{is})\hat{\mu}_{it}(1-\hat{\mu}_{it})}},$$

$\hat{\mu}_{it} = (1+\exp(X_{it}'\boldsymbol{\beta}^{(l)}))^{-1}$; $\hat{p}_{ist}$, an estimate of $E_\xi(Y_{is}Y_{it}) = P(Y_{is}=1, Y_{it}=1)$, given for example in Liang et al. (1992) or Lipsitz et al. (1991), has the form

$$\hat{p}_{ist} = \begin{cases} \dfrac{f_{ist} - \{f_{ist}^2 - 4\widehat{OR}_{st}(\widehat{OR}_{st}-1)\hat{\mu}_{is}\hat{\mu}_{it}\}^{1/2}}{2(\widehat{OR}_{st}-1)}, & \text{if } \widehat{OR}_{st} \neq 1 \\ \widehat{OR}_{st}\hat{\mu}_{is}\hat{\mu}_{it}, & \text{if } \widehat{OR}_{st} = 1, \end{cases}$$

and $f_{ist} = 1 - (1-\widehat{OR}_{st})(\hat{\mu}_{is}+\hat{\mu}_{it})$.

Then we compute the updated estimate $\boldsymbol{\beta}^{(l+1)}$ with:

$$\boldsymbol{\beta}^{(l+1)} = \boldsymbol{\beta}^{(l)} + \Big(\sum_s w_i \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}^{(l)}} [\hat{A}_i^{1/2} \hat{\mathbf{R}}_i^{(l)} \hat{A}_i^{1/2}]^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}^{(l)}}\Big)^{-1} \sum_s w_i \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}^{(l)}} [\hat{A}_i^{1/2} \hat{\mathbf{R}}_i^{(l)} \hat{A}_i^{1/2}]^{-1} (\boldsymbol{y}_i^* - \hat{\boldsymbol{\mu}}_i),$$

where $\partial\hat{\boldsymbol{\mu}}_i/\partial\boldsymbol{\beta}^{(l)} = (\partial\hat{\mu}_{i1}/\partial\boldsymbol{\beta}^{(l)}, \partial\hat{\mu}_{i2}/\partial\boldsymbol{\beta}^{(l)}, \partial\hat{\mu}_{i3}/\partial\boldsymbol{\beta}^{(l)}, \partial\hat{\mu}_{i4}/\partial\boldsymbol{\beta}^{(l)})$, $\partial\hat{\mu}_{it}/\partial\boldsymbol{\beta}^{(l)} = \hat{\mu}_{it}(1-\hat{\mu}_{it})X_{it}$; and $\hat{A}_i = \text{diag}[\hat{\mu}_{it}(1-\hat{\mu}_{it})]$.

After convergence we set $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(l)}$ and $\hat{\mathbf{R}}_i = \mathbf{R}_i^{(l)}$. Then we estimate the total variance, as in section 4.3, with $\hat{\mu}_{it} = (1+\exp(X_{it}'\hat{\boldsymbol{\beta}}))^{-1}$, $\hat{A}_i = \text{diag}[\hat{\mu}_{it}(1-\hat{\mu}_{it})]$, the $\hat{\mathbf{R}}_i$'s are calculated as in eq. 5.9, but using $\hat{\boldsymbol{\beta}}$, and $\hat{V}_i = \hat{A}_i^{1/2}\hat{\mathbf{R}}_i\hat{A}_i^{1/2}$.

For the stratified sampling case (STSI), some of the expressions in section 4.3 can be written in the following way. The naïve estimator of the inner piece of V_{Sam} is:

$$meat(\hat{V}_\pi^*) = \frac{1}{N^2}\sum_{h=1}^{2} \frac{1}{n_h - 1}\Big\{n_h \sum_{k\in s_h} [w_k \frac{\partial \hat{\boldsymbol{\mu}}_k'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_k^{-1} \boldsymbol{r}_k^*]^{\otimes 2} - \Big(\sum_{i\in s_h} w_i \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \boldsymbol{r}_i^*\Big)^{\otimes 2}\Big\},$$

where $\boldsymbol{r}_k^* = (\boldsymbol{y}_k^* - \hat{\boldsymbol{\mu}}_i)$.

The inner piece of the estimated bias, in eq. 4.23, is:

$$meat(Bias[\hat{V}_\pi^*]) = \frac{1}{N^2}\sum_{h=1}^{2}\Bigg\{\underbrace{\sum_{k\in s_h} w_k^2 \frac{\partial \hat{\boldsymbol{\mu}}_k'}{\partial \hat{\boldsymbol{\beta}}}\hat{V}_k^{-1}\left[\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_k^{\mathrm{I}\prime}\end{pmatrix} - \begin{pmatrix}\mathbf{0} & \hat{V}_k^{(\mathrm{oM})}\\ \hat{V}_k^{(\mathrm{oM})\prime} & \hat{V}_k^{(\mathrm{MM})}\end{pmatrix}\right]\hat{V}_k^{-1}\frac{\partial \hat{\boldsymbol{\mu}}_k}{\partial \hat{\boldsymbol{\beta}}}}_{\text{single summation term}}$$
$$\underbrace{-\frac{1}{n_h-1}\sum_{k,i\in s_h}\sum_{i\neq k} w_k w_i \frac{\partial \hat{\boldsymbol{\mu}}_k'}{\partial \hat{\boldsymbol{\beta}}}\hat{V}_k^{-1}\begin{pmatrix}\mathbf{0} & \boldsymbol{r}_k^{\mathrm{o}}\boldsymbol{r}_i^{\mathrm{I}\prime}\\ \boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_i^{\mathrm{o}\prime} & \boldsymbol{r}_k^{\mathrm{I}}\boldsymbol{r}_i^{\mathrm{I}\prime}\end{pmatrix}\hat{V}_i^{-1}\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\boldsymbol{\beta}}}}_{\text{double summation term}}\Bigg\},$$

where $\boldsymbol{r}_k^{\mathrm{o}} = \boldsymbol{y}_k^{\mathrm{o}} - \hat{\boldsymbol{\mu}}_k^{\mathrm{o}}$ and $\boldsymbol{r}_k^{\mathrm{I}} = \boldsymbol{y}_k^{\mathrm{I}} - \hat{\boldsymbol{\mu}}_k^{\mathrm{M}}$. And the estimator of the middle part of $C_{\text{Imp-Sam}}$ is:

$$meat(\hat{C}_{\text{Imp-Sam}}) = \frac{1}{N^2}\sum_{h=1}^{2}\Bigg\{\underbrace{\sum_{i,j\in s_h}\sum_{i\neq j} w_i w_j \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}}\hat{V}_i^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \boldsymbol{r}_i^{\mathrm{I}}\boldsymbol{r}_j^{\mathrm{o}\prime} & \mathbf{0}\end{pmatrix}\hat{V}_j^{-1}\frac{\partial \hat{\boldsymbol{\mu}}_j}{\partial \hat{\boldsymbol{\beta}}}}_{\text{double summation term}}$$
$$\underbrace{-\sum_{s_h} w_i^2 \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}}\hat{V}_i^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \hat{V}_i^{(\mathrm{oM})\prime} & \hat{V}_i^{(\mathrm{MM})}\end{pmatrix}\hat{V}_i^{-1}\frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\boldsymbol{\beta}}}}_{\text{single summation term}}\Bigg\}.$$

And for the cluster sampling case (SIC) some of these expressions are:

$$meat(\hat{V}_\pi^*) = \frac{1}{N^2(n_I-1)}\left\{n_I\sum_{k\in s_I}\left[\sum_{l\in U_k} w_l \frac{\partial \hat{\boldsymbol{\mu}}_l'}{\partial \hat{\boldsymbol{\beta}}}\hat{V}_l^{-1}\boldsymbol{r}_l^*\right]^{\otimes 2} - \left(\sum_{k\in s_I}\sum_{l\in U_k} w_l \frac{\partial \hat{\boldsymbol{\mu}}_l'}{\partial \hat{\boldsymbol{\beta}}}\hat{V}_l^{-1}\boldsymbol{r}_l^*\right)^{\otimes 2}\right\},$$

$$meat(Bias[\hat{V}_\pi^*]) =$$
$$\frac{1}{N^2}\Bigg[\underbrace{\underbrace{\sum_{k\in s_I}\sum_{l\in U_k} w_l^2 \frac{\partial \hat{\boldsymbol{\mu}}_l'}{\partial \hat{\boldsymbol{\beta}}}\hat{V}_l^{-1}\left[\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{r}_l^{\mathrm{I}}\boldsymbol{r}_l^{\mathrm{I}\prime}\end{pmatrix} - \begin{pmatrix}\mathbf{0} & \hat{V}_l^{(\mathrm{oM})}\\ \hat{V}_l^{(\mathrm{oM})\prime} & \hat{V}_l^{(\mathrm{MM})}\end{pmatrix}\right]\hat{V}_l^{-1}\frac{\partial \hat{\boldsymbol{\mu}}_l}{\partial \hat{\boldsymbol{\beta}}}}_{\text{single summation term}}}_{\text{within cluster}}$$
$$\underbrace{+\sum_{k\in s_I}\sum_{l,m\in U_k}\sum_{m\neq l} w_l w_m \frac{\partial \hat{\boldsymbol{\mu}}_l'}{\partial \hat{\boldsymbol{\beta}}}\hat{V}_l^{-1}\begin{pmatrix}\mathbf{0} & \boldsymbol{r}_l^{\mathrm{o}}\boldsymbol{r}_m^{\mathrm{I}\prime}\\ \boldsymbol{r}_l^{\mathrm{I}}\boldsymbol{r}_m^{\mathrm{o}\prime} & \boldsymbol{r}_l^{\mathrm{I}}\boldsymbol{r}_m^{\mathrm{I}\prime}\end{pmatrix}\hat{V}_m^{-1}\frac{\partial \hat{\boldsymbol{\mu}}_m}{\partial \hat{\boldsymbol{\beta}}}}_{\text{double summation}}$$
$$\underbrace{-\frac{1}{n_I-1}\sum_{k,i\in s_I}\sum_{i\neq k}\sum_{l\in U_k}\sum_{m\in U_i} w_l w_m \frac{\partial \hat{\boldsymbol{\mu}}_l'}{\partial \hat{\boldsymbol{\beta}}}\hat{V}_l^{-1}\begin{pmatrix}\mathbf{0} & \boldsymbol{r}_l^{\mathrm{o}}\boldsymbol{r}_m^{\mathrm{I}\prime}\\ \boldsymbol{r}_l^{\mathrm{I}}\boldsymbol{r}_m^{\mathrm{o}\prime} & \boldsymbol{r}_l^{\mathrm{I}}\boldsymbol{r}_m^{\mathrm{I}\prime}\end{pmatrix}\hat{V}_m^{-1}\frac{\partial \hat{\boldsymbol{\mu}}_m}{\partial \hat{\boldsymbol{\beta}}}}_{\text{between cluster (double summation)}}\Bigg], \tag{5.10}$$

and

$$meat(\hat{C}_{\text{Imp-Sam}}) = \frac{1}{N^2}\Bigg\{\underbrace{\sum_{k\in s_I}\sum_{m\in s_I}\sum_{i\in U_k}\sum_{\substack{j\in U_m\\ j\neq i}} w_i w_j \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}}\hat{V}_i^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \boldsymbol{r}_i^{\mathrm{I}}\boldsymbol{r}_j^{\mathrm{o}\prime} & \mathbf{0}\end{pmatrix}\hat{V}_j^{-1}\frac{\partial \hat{\boldsymbol{\mu}}_j}{\partial \hat{\boldsymbol{\beta}}}}_{\text{double summation term}}$$

$$\underbrace{-\sum_{k\in s_I}\sum_{i\in U_k} w_i^2 \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \hat{V}_i^{(\mathrm{oM})'} & \hat{V}_i^{(\mathrm{MM})} \end{pmatrix} \hat{V}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\boldsymbol{\beta}}}}_{\text{single summation term}} \Bigg\}.$$

The alternative variance estimation of section 4.4 takes the following form. The inner part of the imputation variance component is estimated by:

$$\widehat{\mathrm{Var}}_I[U_n^*(\boldsymbol{\beta})] = \frac{1}{N^2}\sum_s w_i^2 \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(s_{\tau r}^2) \end{pmatrix} \hat{V}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\boldsymbol{\beta}}}.$$

The inner part of the sampling variance component is estimated by $(n/r)^2 \hat{V}_{\tau\text{naïve}}$, where, for SRS,

$$\hat{V}_{\tau\text{naïve}} = \frac{1}{N^2(n-1)}\Big[n\sum_{k\in s} w_k^2 \hat{\boldsymbol{z}}_{\tau k}\hat{\boldsymbol{z}}_{\tau k}' - \Big(\sum_{i\in s} w_i \hat{\boldsymbol{z}}_{\tau i}\Big)^{\otimes 2}\Big];$$

for STSI:

$$\hat{V}_{\tau\text{naïve}} = \frac{1}{N^2}\sum_{h=1}^{2} \frac{1}{n_h - 1}\Big[n_h \sum_{k\in s_h} w_k^2 \hat{\boldsymbol{z}}_{\tau k}\hat{\boldsymbol{z}}_{\tau k}' - \Big(\sum_{i\in s_h} w_i \hat{\boldsymbol{z}}_{\tau i}\Big)^{\otimes 2}\Big];$$

and for SIC:

$$\hat{V}_{\tau\text{naïve}} = \frac{1}{N^2(n_I - 1)}\Big[n_I \sum_{k\in s_I}\Big[\sum_{l\in U_k} w_l \hat{\boldsymbol{z}}_{\tau l}\Big]^{\otimes 2} - \Big(\sum_{i\in s_I}\sum_{l\in U_i} w_l \hat{\boldsymbol{z}}_{\tau l}\Big)^{\otimes 2}\Big];$$

where

$$\hat{\boldsymbol{z}}_{\tau i} = \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \begin{pmatrix} \boldsymbol{y}_i^{\mathrm{o}} - \hat{\boldsymbol{\mu}}_i^{\mathrm{o}} \\ \bar{y}_{\tau r} - \hat{\boldsymbol{\mu}}_i^{\mathrm{M}} \end{pmatrix}.$$

5.3 Results

We evaluate the point estimator obtained by Pseudo-GEE with unweighted hot-deck imputed responses along two traits, relative bias and variance (MSE). We also study how the variance estimator(s) developed in section 4.3 perform for the chosen sample sizes and probabilities of missingness.

Our simulations are programmed in the R software package, as documented in R Development Core Team (2008). All simulations are run on a UNIX machine with 24 CPUs.

5.3.1 Point Estimation

Based on the 1,000 estimated values of $\hat{\boldsymbol{\beta}}$, for each combination of sample size and p_m, we estimate the relative bias of the estimator $\hat{\boldsymbol{\beta}}$ with

$$RB(\hat{\boldsymbol{\beta}}) = \Big(\frac{1}{1000}\sum_{k=1}^{1000} \hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta}\Big)/\boldsymbol{\beta},$$

where $\hat{\boldsymbol{\beta}}^{(k)}$ is the estimated value from the k-th sample, and the division is carried out term by term. The summary of the relative biases is presented in Tables 5.1–5.6.

For all sampling schemes considered and for either continuous or binary response, the biggest relative bias (in absolute value) is about 5%, which occurs with the smallest sample size of 240. This case corresponds to having only around 10 selected subjects per cell. For the rest of cases the biggest relative bias is about 3%. And for sample sizes of 720 (about 30 per cell) and above, the maximum relative bias is bounded by around 2%, for all missing fractions considered.

A good feature of the proposed estimator is that, as the sample size increases, its relative bias tends to decrease; although this is not in all cases a totally monotone pattern. Also, the bias of the estimator does not seem to be influenced by the missing percentage; at least for the MAR case and missing fractions considered. Given a sample size, a higher missing rate does not necessarily carry about a higher bias. This is a good characteristic because once we have a sample of a fixed size, we do not like the performance of the estimator to depend on how many respondents there are; as long as the missing fraction is within the limits examined here, and the nonresponse satisfies a MAR mechanism.

Estimators of the regression coefficients under stratified sampling seem to perform a little better than those under simple random sampling for the continuous response. Whereas for the binary response the opposite holds. This indicates that the stratification used (based on age) is not as efficient for the latter as it is for the former case.

Table 5.1: Rel. Bias of $\hat{\boldsymbol{\beta}}$, in %, Continuous Response, SRS

n	$\bar{n}_c$	p_m	$\bar{r}_c$	β_0	β_1	β_2	β_3	β_4
240	10	0.25	7.50	0.21	0.40	0.48	1.81	1.08
		0.20	8.00	0.01	0.20	0.35	−0.51	−2.93
		0.15	8.50	−0.09	0.02	0.09	−0.48	−4.21
		0.10	9.00	−0.23	−0.13	−0.09	−0.13	−4.54
		0.05	9.50	−0.19	−0.13	−0.07	−0.18	−3.70
480	20	0.25	15.00	−0.11	−0.09	0.00	0.51	−0.05
		0.20	16.00	−0.20	−0.11	0.03	0.51	−3.03
		0.15	17.00	−0.24	−0.18	−0.07	0.64	−2.51
		0.10	18.00	−0.32	−0.33	−0.23	0.20	−2.14
		0.05	19.00	−0.34	−0.35	−0.26	0.10	−2.82
720	30	0.25	22.50	−0.17	−0.43	−0.58	−0.26	1.35
		0.20	24.00	0.25	0.32	0.32	−0.58	−0.75
		0.15	25.50	0.18	0.20	0.18	−0.48	−0.39
		0.10	27.00	0.26	0.30	0.24	−0.33	−0.54
		0.05	28.50	0.21	0.19	0.11	−0.69	−0.46
960	40	0.25	30.00	0.12	0.18	0.20	−0.51	−0.33
		0.20	32.00	0.13	0.15	0.13	−0.43	0.48
		0.15	34.00	0.15	0.13	0.12	−0.70	0.47
		0.10	36.00	0.12	0.08	0.05	−0.89	0.26
		0.05	38.00	0.12	0.11	0.09	−0.55	0.32
1200	50	0.25	37.50	−0.08	−0.16	−0.15	−0.80	−1.33
		0.20	40.00	0.08	0.09	0.18	−0.49	0.06
		0.15	42.50	0.01	0.01	0.09	−0.17	0.23
		0.10	45.00	−0.07	−0.16	−0.12	−0.53	0.01
		0.05	47.50	0.01	−0.01	0.05	−0.36	−0.05

n = sample size

$\bar{n}_c$ = average cell sample size

p_m = probability of non-response

$\bar{r}_c$ = average number of respondents per cell

Table 5.2: Rel. Bias of $\hat{\boldsymbol{\beta}}$, in %, Continuous Response, STSI

n	$\bar{n}_c$	p_m	$\bar{r}_c$	β_0	β_1	β_2	β_3	β_4
240	10	0.25	7.50	−0.35	−0.57	−0.56	1.36	1.08
		0.20	8.00	0.27	0.40	0.55	0.26	1.15
		0.15	8.50	0.40	0.54	0.63	0.57	1.53
		0.10	9.00	0.35	0.40	0.45	0.41	2.26
		0.05	9.50	0.29	0.28	0.29	0.19	2.32
480	20	0.25	15.00	−0.32	−0.78	−0.97	0.22	2.51
		0.20	16.00	0.16	0.21	0.34	−0.80	1.68
		0.15	17.00	−0.01	−0.02	0.13	−0.55	1.66
		0.10	18.00	0.02	−0.06	−0.01	−0.96	1.69
		0.05	19.00	0.05	0.03	0.11	−0.77	1.31
720	30	0.25	22.50	0.29	0.30	0.29	0.66	2.51
		0.20	24.00	0.02	−0.12	−0.17	0.21	0.10
		0.15	25.50	0.02	−0.13	−0.18	0.51	0.58
		0.10	27.00	0.00	−0.18	−0.23	0.15	0.83
		0.05	28.50	−0.10	−0.33	−0.41	0.39	1.34
960	40	0.25	30.00	0.01	0.03	0.04	−0.29	−0.48
		0.20	32.00	0.03	0.04	0.02	0.32	0.80
		0.15	34.00	0.03	0.07	0.05	0.42	0.23
		0.10	36.00	0.02	−0.01	−0.07	0.18	0.68
		0.05	38.00	−0.01	−0.07	−0.14	−0.05	0.50
1200	50	0.25	37.50	−0.06	−0.14	−0.16	0.09	−0.40
		0.20	40.00	0.14	0.33	0.47	0.57	0.64
		0.15	42.50	0.21	0.40	0.54	0.65	0.80
		0.10	45.00	0.07	0.18	0.28	0.52	0.60
		0.05	47.50	0.04	0.12	0.21	0.63	0.71

Table 5.3: Rel. Bias of $\hat{\boldsymbol{\beta}}$, in %, Continuous Response, SIC

n_I	$E_\pi(n)$	$\bar{n}_c$	p_m	$\bar{r}_c$	β_0	β_1	β_2	β_3	β_4
36	240	10	0.25	7.50	−0.23	−0.40	−0.54	−1.13	1.94
			0.20	8.00	−0.34	−0.52	−0.63	0.57	2.28
			0.15	8.50	−0.45	−0.70	−0.85	−0.01	2.32
			0.10	9.00	−0.35	−0.43	−0.47	0.00	1.10
			0.05	9.50	−0.27	−0.27	−0.27	0.66	1.65
72	480	20	0.25	15.00	0.19	0.35	0.49	−1.09	1.79
			0.20	16.00	0.23	0.22	0.08	−0.90	0.19
			0.15	17.00	0.21	0.21	0.07	−0.69	0.64
			0.10	18.00	0.20	0.21	0.07	−0.39	−0.08
			0.05	19.00	0.13	0.07	−0.12	−0.62	−0.36
108	720	30	0.25	22.50	0.10	0.15	0.19	1.08	0.70
			0.20	24.00	0.02	−0.05	−0.05	−0.29	1.29
			0.15	25.50	0.01	0.02	0.05	0.23	1.38
			0.10	27.00	0.07	0.12	0.18	0.15	0.81
			0.05	28.50	0.04	0.08	0.10	0.01	0.32
144	960	40	0.25	30.00	−0.07	0.13	0.31	0.49	−0.46
			0.20	32.00	−0.11	0.11	0.25	0.80	−1.86
			0.15	34.00	−0.03	0.17	0.31	0.26	−1.53
			0.10	36.00	0.01	0.24	0.41	0.22	−1.41
			0.05	38.00	−0.01	0.22	0.41	0.30	−1.31
180	1200	50	0.25	37.50	−0.02	−0.08	−0.15	0.12	−0.97
			0.20	40.00	0.08	−0.13	−0.36	0.20	0.55
			0.15	42.50	0.17	0.01	−0.20	0.46	0.76
			0.10	45.00	0.13	−0.05	−0.27	0.74	0.67
			0.05	47.50	0.09	−0.09	−0.32	0.64	0.03

Table 5.4: Rel. Bias of $\hat{\boldsymbol{\beta}}$, in %, Binary Response, SRS

n	$\bar{n}_c$	p_m	$\bar{r}_c$	β_0	β_1	β_2	β_3	β_4
240	10	0.25	7.50	1.29	1.37	1.55	0.89	−0.59
		0.20	8.00	0.33	−0.19	−0.77	−0.66	1.23
		0.15	8.50	0.35	−0.05	−0.52	−0.97	0.15
		0.10	9.00	0.65	0.23	−0.20	−1.07	0.64
		0.05	9.50	0.53	0.22	−0.12	−0.92	−0.19
480	20	0.25	15.00	0.81	0.53	0.52	−0.60	−1.06
		0.20	16.00	−0.10	−0.01	−0.15	0.42	−1.52
		0.15	17.00	0.06	0.23	0.16	0.79	−1.62
		0.10	18.00	0.09	0.24	0.17	0.69	−1.26
		0.05	19.00	−0.10	0.03	−0.10	0.56	−1.44
720	30	0.25	22.50	0.09	0.06	−0.05	0.14	0.52
		0.20	24.00	0.16	0.13	0.17	−1.14	0.33
		0.15	25.50	0.04	0.10	0.12	−0.70	−0.50
		0.10	27.00	−0.22	−0.15	−0.17	−0.44	−0.20
		0.05	28.50	−0.06	0.01	0.04	−0.80	−0.45
960	40	0.25	30.00	0.03	−0.14	−0.37	0.36	0.59
		0.20	32.00	0.47	0.37	0.23	0.78	−1.05
		0.15	34.00	0.38	0.30	0.15	0.66	−1.08
		0.10	36.00	0.28	0.24	0.10	0.77	−1.27
		0.05	38.00	0.27	0.21	0.09	0.48	−1.04
1200	50	0.25	37.50	0.09	−0.02	−0.16	0.10	0.56
		0.20	40.00	−0.06	−0.04	−0.14	0.19	−0.84
		0.15	42.50	0.00	0.01	−0.07	−0.02	−0.30
		0.10	45.00	0.12	0.15	0.14	−0.03	−0.67
		0.05	47.50	−0.01	0.01	−0.04	0.05	−0.70

Table 5.5: Rel. Bias of $\hat{\boldsymbol{\beta}}$, in %, Binary Response, STSI

n	$\bar{n}_c$	p_m	$\bar{r}_c$	β_0	β_1	β_2	β_3	β_4
240	10	0.25	7.50	2.01	1.51	1.53	0.12	3.57
		0.20	8.00	1.71	1.43	1.59	0.48	2.96
		0.15	8.50	1.41	1.21	1.34	0.71	2.85
		0.10	9.00	1.65	1.28	1.37	0.17	3.48
		0.05	9.50	1.54	1.20	1.26	0.29	3.61
480	20	0.25	15.00	0.43	0.19	0.10	−1.12	−0.23
		0.20	16.00	−0.12	−0.40	−0.72	0.54	2.65
		0.15	17.00	−0.37	−0.65	−1.04	0.58	2.32
		0.10	18.00	−0.16	−0.45	−0.77	0.45	2.35
		0.05	19.00	−0.19	−0.40	−0.67	0.49	1.43
720	30	0.25	22.50	−0.23	−0.04	−0.06	−0.73	−2.20
		0.20	24.00	−0.01	0.03	−0.15	−0.31	−1.48
		0.15	25.50	0.09	0.08	−0.10	−0.59	−1.61
		0.10	27.00	0.21	0.14	−0.03	−0.89	−0.83
		0.05	28.50	0.39	0.31	0.17	−0.94	−1.23
960	40	0.25	30.00	0.34	0.48	0.55	0.43	−1.40
		0.20	32.00	−0.29	−0.21	−0.33	0.79	−1.20
		0.15	34.00	−0.43	−0.28	−0.38	0.72	−1.34
		0.10	36.00	−0.57	−0.40	−0.54	0.88	−1.47
		0.05	38.00	−0.32	−0.19	−0.25	0.62	−0.99
1200	50	0.25	37.50	0.28	0.25	0.31	0.38	2.11
		0.20	40.00	0.05	0.11	0.12	0.02	1.53
		0.15	42.50	−0.03	−0.02	−0.01	−0.38	1.32
		0.10	45.00	−0.07	−0.06	−0.07	−0.15	1.46
		0.05	47.50	−0.13	−0.10	−0.11	−0.21	1.27

Table 5.6: Rel. Bias of $\hat{\boldsymbol{\beta}}$, in %, Binary Response, SIC

n_I	$E_\pi(n)$	$\bar{n}_c$	p_m	$\bar{r}_c$	β_0	β_1	β_2	β_3	β_4
36	240	10	0.25	7.50	1.16	0.79	0.42	−1.33	−1.71
			0.20	8.00	0.72	0.78	0.67	0.91	−1.24
			0.15	8.50	0.72	0.72	0.56	0.77	−1.76
			0.10	9.00	0.50	0.61	0.51	1.14	−1.36
			0.05	9.50	0.59	0.62	0.54	0.23	−1.64
72	480	20	0.25	15.00	0.73	0.63	0.68	−0.78	0.39
			0.20	16.00	0.77	0.47	0.19	0.33	−0.56
			0.15	17.00	0.56	0.34	0.10	−0.20	−1.06
			0.10	18.00	0.53	0.25	0.02	−0.25	−0.77
			0.05	19.00	0.61	0.40	0.28	−0.50	−0.92
108	720	30	0.25	22.50	0.81	0.62	0.48	0.51	0.99
			0.20	24.00	0.57	0.41	0.32	−0.56	−0.29
			0.15	25.50	0.49	0.36	0.28	−0.71	−0.32
			0.10	27.00	0.20	0.14	0.05	−0.48	−0.89
			0.05	28.50	0.25	0.23	0.20	−0.38	−1.13
144	960	40	0.25	30.00	−0.10	−0.02	0.00	−0.63	−1.86
			0.20	32.00	0.03	0.10	0.18	−0.47	−1.94
			0.15	34.00	0.09	0.16	0.25	−0.20	−1.72
			0.10	36.00	0.15	0.26	0.36	−0.24	−2.22
			0.05	38.00	0.09	0.16	0.25	−0.28	−1.16
180	1200	50	0.25	37.50	0.00	−0.15	−0.36	−0.21	0.59
			0.20	40.00	0.27	0.19	0.11	−1.15	−1.65
			0.15	42.50	0.27	0.14	0.06	−1.66	−1.80
			0.10	45.00	0.09	0.02	−0.05	−1.23	−1.84
			0.05	47.50	0.10	0.01	−0.05	−1.36	−1.74

From the 1,000 simulated values we can also find (approximately) the true variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, based on the following formula:

$$var(\hat{\boldsymbol{\beta}}) = \frac{1}{1000}\sum_{k=1}^{1000}(\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}^{(k)} - \boldsymbol{\beta})'.$$

For the continuous response and cluster sampling [3], the summary of these "true" variances can be found in Tables 5.7–5.9, for some of the selected sample sizes. These values not only help us address the performance of the estimator $\hat{\boldsymbol{\beta}}$ itself, but also give us quantities to compare our variance estimators to later on.

It is clear that as the sample size increases, or as the missing percentage reduces, the variance of $\hat{\boldsymbol{\beta}}$ decreases. However, it is noteworthy that the variance does not necessarily decrease as the number of *respondents* increases. In some situations the variance goes up even when the number of actual respondents goes up, if the missing *fraction* increases.

Additionally we should point out that some of the covariances of $\hat{\boldsymbol{\beta}}$ are remarkably close to zero. This suggests that evaluating the performance of the variance-covariance estimators with a measure of relative bias that divided by such covariance terms would not be a good idea. That is why, in the next section, for the covariance terms we use a relative bias that does not divide by these terms.

[3] The other results are omitted since they convey the same message.

Table 5.7: MSE of $\hat{\boldsymbol{\beta}}$, Continuous Response, SIC, $n = 240$ (average cell sample size $\bar{n}_c = 10$)

p_m	$\bar{r}_c$	Variance-covaraince matrix				
0.25	7.50	0.452616				
		−0.127753	0.042866			
		0.008502	−0.003077	0.000234		
		−0.002285	−0.000102	0.000010	0.000620	
		−0.022319	−0.001458	0.000142	0.000415	0.049948
0.20	8.00	0.415931				
		−0.116098	0.038737			
		0.007905	−0.002842	0.000221		
		−0.001724	−0.000038	0.000007	0.000580	
		−0.029194	0.001181	−0.000061	−0.000237	0.047229
0.15	8.50	0.396637				
		−0.110821	0.036933			
		0.007522	−0.002698	0.000209		
		−0.001363	−0.000123	0.000014	0.000535	
		−0.026925	0.001227	−0.000072	−0.000287	0.044563
0.10	9.00	0.367805				
		−0.101666	0.033686			
		0.006854	−0.002453	0.000190		
		−0.001275	−0.000036	0.000005	0.000467	
		−0.024167	0.000532	−0.000020	−0.000278	0.042471
0.05	9.50	0.346779				
		−0.095421	0.031512			
		0.006389	−0.002282	0.000176		
		−0.001014	−0.000025	0.000002	0.000385	
		−0.023547	0.000592	−0.000021	−0.000251	0.039993

Table 5.8: MSE of $\hat{\boldsymbol{\beta}}$, Continuous Response, SIC, $n = 720$ (average cell sample size $\bar{n}_c = 30$)

p_m	$\bar{r}_c$	Variance-covaraince matrix				
0.25	7.50	0.154118				
		−0.043590	0.014351			
		0.002950	−0.001035	0.000079		
		−0.000757	0.000010	0.000000	0.000198	
		−0.009518	0.000446	−0.000024	0.000045	0.016340
0.20	8.00	0.131352				
		−0.036755	0.012343			
		0.002475	−0.000901	0.000070		
		−0.000761	0.000042	−0.000001	0.000172	
		−0.007499	−0.000126	0.000010	−0.000003	0.015429
0.15	8.50	0.125305				
		−0.035264	0.011931			
		0.002375	−0.000872	0.000068		
		−0.000681	0.000024	0.000000	0.000157	
		−0.006969	−0.000216	0.000025	0.000032	0.014331
0.10	9.00	0.112994				
		−0.031574	0.010727			
		0.002105	−0.000781	0.000060		
		−0.000637	0.000017	0.000001	0.000145	
		−0.005722	−0.000463	0.000040	0.000032	0.013493
0.05	9.50	0.103234				
		−0.028840	0.009827			
		0.001919	−0.000715	0.000055		
		−0.000543	0.000015	0.000001	0.000131	
		−0.005703	−0.000286	0.000028	0.000001	0.012898

Table 5.9: MSE of $\hat{\boldsymbol{\beta}}$, Continuous Response, SIC, $n = 1200$ (average cell sample size $\bar{n}_c = 50$)

p_m	$\bar{r}_c$	Variance-covaraince matrix				
0.25	7.50	0.084052				
		−0.023795	0.007939			
		0.001618	−0.000578	0.000045		
		−0.000311	−0.000020	0.000002	0.000111	
		−0.005800	0.000280	−0.000015	−0.000030	0.009606
0.20	8.00	0.073693				
		−0.020830	0.007041			
		0.001414	−0.000515	0.000040		
		−0.000461	0.000030	−0.000001	0.000100	
		−0.004342	−0.000078	0.000011	0.000003	0.008630
0.15	8.50	0.071043				
		−0.020002	0.006712			
		0.001358	−0.000490	0.000038		
		−0.000443	0.000035	−0.000002	0.000091	
		−0.004385	−0.000063	0.000009	0.000029	0.008575
0.10	9.00	0.065592				
		−0.018292	0.006104			
		0.001237	−0.000445	0.000034		
		−0.000355	0.000021	−0.000001	0.000081	
		−0.005019	0.000173	−0.000008	0.000019	0.008340
0.05	9.50	0.060568				
		−0.016860	0.005653			
		0.001134	−0.000411	0.000032		
		−0.000297	0.000016	−0.000001	0.000068	
		−0.004519	0.000103	−0.000001	0.000022	0.007754

5.3.2 Variance Estimation

For variance estimation, there are several ways to carry out the task. This is so because, in section 4.3, we have decomposed the variance of $\hat{\boldsymbol{\beta}}$, and its estimation, into several pieces. The first and easiest option is to add up only $\hat{V}_{\text{Imp}}$ and $\hat{V}^*_{\pi}$; i.e. an estimator of the imputation variance and the naïve sampling component. For estimation of finite population totals, Särndal (1992) argues that the mixed (imputation-sampling) component can be ignored "if the expected imputation error is zero or negligible under the response mechanism, conditional on the realized sample." He also points out that when imputing a "predicted value plus residual", which hot-deck can be thought of, the bias of the naïve sampling component is negligible (for estimation of totals). We call this estimator V1:

$$V1 = \hat{V}_{\text{Imp}} + \hat{V}^*_{\pi}.$$

The second option is to add $\hat{V}_{\text{Imp}}$ and $\hat{V}^*_{\pi}$, as in V1, but also the terms with the single summations in eq. 4.24 (i.e. ignoring the term in the double summation). This is an attempt to correct for the mixed imputation-sampling term, but without the hassle of calculating the double summation in $\hat{C}_{\text{Imp-Sam}}$, which is computationally time consuming. We call this estimator V2:

$$\begin{aligned} V2 = \hat{V}_{\text{Imp}} + \hat{V}^*_{\pi} + \Big[\frac{1}{N}\sum_s w_i \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\boldsymbol{\beta}}}\Big]^{-1} \times \\ \Big[\frac{-1}{N^2}\sum_s w_i^2 \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \hat{V}_i^{(\text{oM})'} & \hat{V}_i^{(\text{MM})} \end{pmatrix} \hat{V}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\boldsymbol{\beta}}}\Big] \times \\ \Big[\frac{1}{N}\sum_s w_i \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\boldsymbol{\beta}}}\Big]^{-1} \\ +\text{transpose of last "sandwich" term.} \end{aligned}$$

Another way to estimate the variance of $\hat{\boldsymbol{\beta}}$ is to add $\hat{V}_{\text{Imp}}$ and $\hat{V}^*_{\pi}$, and also the full mixed term $\hat{C}_{\text{Imp-Sam}} + \hat{C}'_{\text{Imp-Sam}}$. For linear estimators, Brick et al. (2004) show some examples in which the contribution of this mixed term to the total variance is substantial, and then they recommend not ignoring it. This estimator is called V3 here:

$$V3 = \hat{V}_{\text{Imp}} + \hat{V}^*_{\pi} + \hat{C}_{\text{Imp-Sam}} + \hat{C}'_{\text{Imp-Sam}}.$$

There are two ways in which we can correct for the bias of the naïve sampling component estimator, $\hat{V}^*_{\pi}$. The "fast and easy" way is to subtract just the term in the single summation in expression 4.23, and ignore the double summation term. We call this estimator V4:

$$\begin{aligned} V4 = \hat{V}_{\text{Imp}} + \hat{V}^*_{\pi} - \Big[\frac{1}{N}\sum_s w_i \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}} V_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\boldsymbol{\beta}}}\Big]^{-1} \times \\ \Big[\frac{1}{N^2}\sum_{k\in s} w_k^2 \frac{\partial \hat{\boldsymbol{\mu}}_k'}{\partial \hat{\boldsymbol{\beta}}} \hat{V}_k^{-1} \Big[\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{r}_k^{\text{I}} \boldsymbol{r}_k^{\text{I}\prime} \end{pmatrix} - \begin{pmatrix} \mathbf{0} & \hat{V}_k^{(\text{oM})} \\ \hat{V}_k^{(\text{oM})'} & \hat{V}_k^{(\text{MM})} \end{pmatrix}\Big] \hat{V}_k^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_k}{\partial \hat{\boldsymbol{\beta}}}\Big] \times \end{aligned}$$

$$\Big[\frac{1}{N}\sum_s w_i \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \hat{\boldsymbol{\beta}}} V_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \hat{\boldsymbol{\beta}}}\Big]^{-1}.$$

The other way is to subtract the full bias correction term, including the computationally burdensome double summation piece. This gives V5:

$$\begin{aligned} V5 =& \hat{V}_{\text{Imp}} + \hat{V}_\pi^* - \widehat{\text{Bias}}(\hat{V}_\pi^*) \\ =& \hat{V}_{\text{Imp}} + \hat{V}_{\text{Sam}}, \end{aligned}$$

where $\widehat{\text{Bias}}(\hat{V}_\pi^*)$ is given by expression 4.23.

The most "complete" variance estimator, V6, is the one that adds/subtracts all the components, and every piece in each estimated component:

$$\begin{aligned} V6 =& \hat{V}_{\text{Imp}} + \hat{V}_\pi^* - \widehat{\text{Bias}}(\hat{V}_\pi^*) + \hat{C}_{\text{Imp-Sam}} + \hat{C}'_{\text{Imp-Sam}} \\ =& \hat{V}_{\text{Imp}} + \hat{V}_{\text{Sam}} + \hat{C}_{\text{Imp-Sam}} + \hat{C}'_{\text{Imp-Sam}}. \end{aligned}$$

Finally, we additionally study the suitability of an estimator, V24, which can be considered a "mixture" between V2 and V4. This estimator adds up $\hat{V}_{\text{Imp}}$ and $\hat{V}_\pi^*$, but also attempts to correct for the bias of $\hat{V}_\pi^*$ and for the mixed term, but only using the single summation terms of $\widehat{\text{Bias}}(\hat{V}_\pi^*)$ and $C_{\text{Imp-Sam}}$. In other words, V24 aims to be as complete as V6, but only through single summation terms and avoids any double summation term.

Due to the form that the bias correction term takes for the SIC sampling (equations 5.4 and 5.10), we can think of two additional variance estimators in this case. One that adds $\hat{V}_{\text{Imp}}$ and $\hat{V}_\pi^*$, and also tries to correct the naïve sampling component term but only by adding the "within cluster" term of the bias correction; this variance estimator is called V4A:

$$V4A = \hat{V}_{\text{Imp}} + \hat{V}_\pi^* + \text{within cluster bias correction term}.$$

This estimator contains the single summation term of the bias correction and the double summation term that adds up pair of subjects only when they belong to the same cluster. The other possibility, called V4B, adds $\hat{V}_{\text{Imp}}$ and $\hat{V}_\pi^*$, the single summation term of the bias correction, and the double summation term that adds up pair of subjects only when they belong to different clusters:

$$\begin{aligned} V4B = \hat{V}_{\text{Imp}} + \hat{V}_\pi^* &+ \text{single summation term of bias correction} \\ &+ \text{between cluster bias correction term}. \end{aligned}$$

In order to evaluate the performance of each of the nine variance-covariance estimators, we calculate for each of them the relative bias as estimator of V_{Tot}. Since we do not know V_{Tot}, we use our best estimate of it, $var(\hat{\boldsymbol{\beta}})$, as the target. For the i-th diagonal term (variance of $\hat{\beta}_i$) we use:

$$RB(\text{Vj}_i) = \Big(\frac{1}{1000}\sum_{k=1}^{1000} \text{Vj}_i^{(k)} - var(\hat{\boldsymbol{\beta}})_i\Big) / var(\hat{\boldsymbol{\beta}})_i,$$

where Vj is any of V1, V2, V3, V4, V5, V6, V24, V4A, or V4B; $\mathrm{Vj}^{(k)}$ is the calculated value of Vj from the k-th simulation sample, and $\mathrm{Vj}_i^{(k)}$ is its i-th diagonal term; and $var(\hat{\boldsymbol{\beta}})_i$ is the i-th diagonal term of $var(\hat{\boldsymbol{\beta}})$. And for the i,l-th term (i.e. the covariance of $\hat{\beta}_i$ and $\hat{\beta}_l$) we use:

$$RB(\mathrm{Vj}_{il}) = \frac{\frac{1}{1000}\sum_{k=1}^{1000}\mathrm{Vj}_{il}^{(k)} - var(\hat{\boldsymbol{\beta}})_{il}}{\sqrt{var(\hat{\boldsymbol{\beta}})_i}\sqrt{var(\hat{\boldsymbol{\beta}})_l}},$$

where $\mathrm{Vj}_{il}^{(k)}$ is its i,l-th term of $\mathrm{Vj}^{(k)}$; and $var(\hat{\boldsymbol{\beta}})_{il}$ is the i,l-th entry of $var(\hat{\boldsymbol{\beta}})$. In other words, for the covariance terms it is like we are evaluating the absolute bias of the estimator of the corresponding correlation coefficient. This makes more sense than using $var(\hat{\boldsymbol{\beta}})_{il}$ as denominator because some of these terms may be too close to zero.

The relative bias of the variance estimates show that the estimators V2 and V24 perform consistently (much) worse than the others, for both continuous and binary response, all sampling schemes considered, all sample sizes, and all missing percentages. This indicates that the mixed (sampling-imputation) component, if included must be included fully; i.e. if one includes the terms corresponding to the single summations in $\hat{C}_{\text{Imp-Sam}}$, one should also include the terms corresponding to the double summations. Thus we omit the results for these two variance estimators here.

Tables 5.10–5.12 show the relative bias of all the variance estimators for three cases: binary response with SRS and $n = 240$ or with STSI and $n = 1,200$, and for continuous response with SIC and $n = 720$. Results for other cases are alike [4].

For the element sampling cases[5], we notice that V4, V5, and V6 do invariably better than V1 and V3. This is an indication that *some* correction for bias of the naïve sampling variance estimator is crucial. Additionally, V4, V5, and V6 are very close among themselves, and V1 and V3 are close between themselves.

The fact that V1 and V3 are generally close to each other means that the inclusion of the mixed imputation-sampling component is irrelevant. This is further confirmed by the closeness of the estimators V5 and V6. This is very good news as the computation of this term involves a double summation over the sample and is computationally time consuming. And, as we observed earlier, including just the single summation term is disastrous.

As we said, a correction of the naïve sampling component estimator must be incorporated. Nonetheless, the fact that V4 is so close to V5 (and to V6) reveals that it is not necessary to include the whole bias correction component; the single summation term of it suffices. This is also good news; the computation of the double summation part of the bias correction takes a long time. Although V5 is *in a few cases* as much as 1% less biased than V4, in our opinion it is worthless to go through the hassle of calculating the extra term required for that little win. A similar case holds for the most "complete" variance estimator V6; which takes even longer and is harder to calculate than V5.

[4] We also omit the results for 0.15 missing probability, for better fitting of the tables on the pages.

[5] SRS and STSI.

Hence we conclude that, all things being considered, V4 is the best estimator of the variance of $\hat{\boldsymbol{\beta}}$. This is, for the element sampling cases, an estimator of the total variance of $\hat{\boldsymbol{\beta}}$ which includes just the imputation component, the naïve sampling component, and the single summation term of the bias correction, is the best strategy.

For SIC sampling the situation is analogous, but somewhat different. Here we find that V4 and V4B are better than V3, V4A, V5, and V6; that is, the relative biases or the former are consistently smaller than those of the latter. This is particularly true for the larger missing fractions. V4 and V4B are also generally better than V1; except for a couple of cases where they perform similarly. This denotes that the imputation variance together with the naïve sampling variance component are in most cases not enough.

Thus we have that for the SIC sampling considered, besides the imputation variance component and naïve sampling component (which should always be included and are fast and easy to compute), the inclusion of the mixed sampling-imputation component is not necessary, and may even be detrimental. Neither the single summation term nor the double summation term of this component should be added. This is so because V3 and V6 (and V2 and V24) do not perform as well.

The correction for the bias of the naïve sampling component, on the other hand, is a different story. Although V5 does not perform as well as V4 or V4B, which means that including the whole bias correction term is not a good idea, the results show that *some* correction for this bias *is* necessary. Since V4A does not do as well as V4 or V4B, we have that including the "within cluster" part of the bias correction is not a good idea either.

Given that the variance estimators V4 and V4B are the ones that perform generally the best, we have that one should either include only the single summation term of the bias correction or include it together with the "between cluster" part of the bias correction.

Albeit one may be inclined not to include the between cluster bias correction, for it is a double summation term, we do not recommend doing so because we found that V4B is *many times* up to 1% less biased than V4. This is not such a computationally intensive task as it would be in the element sampling case. Here, this term (between clusters) adds up pairs of elements *only* when they belong to different clusters.

So we conclude that, for the SIC sampling case, the best variance estimation strategy is to add up the imputation sampling component, the naïve sampling component, the single summation term of the bias correction, and the between cluster bias correction part.

Table 5.10: Rel. Bias of Var. Estimators, in %, Binary Response, SRS, $n = 240$

	25% missing					20% missing					10% missing					5% missing				
V1	−4					6					2					3				
	4	−5				−7	6				−2	0				−3	2			
	−4	5	−6			7	−6	6			1	0	−1			2	−1	1		
	−3	5	−5	−6		−2	2	−2	−1		−3	3	−3	−1		−3	2	−1	0	
	5	−2	2	0	−18	5	1	−2	−3	−14	1	3	−4	−3	−9	−0	3	−3	−3	−6
V3	−3					6					1					3				
	3	−5				−7	6				−1	0				−3	2			
	−3	5	−5			8	−6	5			1	1	−1			2	−1	1		
	−3	5	−5	−6		−2	2	−2	−1		−3	3	−3	−1		−3	2	−1	0	
	6	−2	2	0	−22	5	1	−2	−3	−17	1	3	−4	−3	−10	−0	3	−3	−3	−7
V4	−7					3					−0					2				
	8	−10				−3	2				1	−2				−2	1			
	−8	10	−11			3	−2	1			−1	3	−4			1	−0	−0		
	−2	4	−4	−11		−1	2	−2	−6		−3	3	−3	−3		−3	2	−1	−1	
	4	−2	2	0	−8	3	1	−2	−4	−5	−1	3	−4	−3	−4	−1	3	−3	−3	−4
V5	−7					3					−0					2				
	8	−10				−3	2				1	−2				−2	1			
	−8	10	−10			3	−2	1			−1	3	−4			1	−0	−0		
	−2	4	−4	−11		−1	2	−2	−6		−3	3	−3	−3		−3	2	−1	−1	
	3	−2	2	0	−8	3	1	−2	−4	−5	−1	3	−4	−3	−4	−1	3	−3	−3	−4
V6	−7					3					−0					2				
	7	−9				−3	2				1	−2				−1	1			
	−7	9	−10			4	−2	1			−1	3	−4			1	−0	−1		
	−2	4	−4	−11		−1	1	−2	−6		−3	3	−3	−3		−3	2	−1	−1	
	4	−2	2	0	−11	3	1	−2	−3	−7	−0	3	−4	−3	−5	−1	3	−3	−3	−4

Table 5.11: Rel. Bias of Var. Estimators, in %, Binary Response, STSI, $n = 1,200$

	25% missing					20% missing					10% missing					5% missing				
	2					4					6					8				
	−4	4				−5	4				−8	8				−9	10			
V1	4	−4	4			4	−3	1			8	−7	6			10	−9	8		
	−3	4	−4	2		−6	6	−5	1		−5	5	−5	6		−4	4	−5	1	
	4	−1	1	−0	−13	4	−2	2	−0	−10	5	−3	2	0	−8	4	−3	1	3	−4
	3					4					5					8				
	−4	5				−5	4				−8	8				−9	10			
V3	5	−5	5			5	−3	2			8	−7	6			9	−9	8		
	−3	4	−4	2		−7	6	−5	1		−5	5	−5	6		−4	4	−4	1	
	4	−1	1	−1	−16	4	−2	2	−0	−13	5	−3	2	0	−9	4	−3	2	3	−5
	−1					2					4					7				
	−0	−0				−1	0				−6	6				−8	8			
V4	1	0	−1			1	1	−3			6	−5	3			8	−8	6		
	−2	3	−4	−2		−6	6	−5	−3		−5	5	−5	4		−4	4	−4	0	
	2	−1	1	−1	−1	2	−2	2	−0	−1	4	−3	2	0	−3	3	−3	1	3	−1
	−1					2					4					7				
	−0	−0				−1	0				−6	6				−8	8			
V5	1	0	−1			1	1	−3			6	−5	3			8	−8	6		
	−2	3	−4	−2		−6	6	−5	−3		−5	5	−5	4		−4	4	−4	0	
	2	−1	1	−1	−1	2	−2	2	−0	−1	4	−3	2	0	−3	3	−3	1	3	−1
	−0					2					4					7				
	−1	1				−2	1				−5	6				−8	8			
V6	2	−1	0			2	0	−2			6	−5	3			8	−7	6		
	−2	3	−4	−2		−6	6	−5	−3		−5	5	−5	4		−4	4	−4	0	
	2	−1	1	−1	−5	2	−2	2	−0	−3	4	−3	2	0	−3	3	−3	2	3	−1

Table 5.12: Rel. Bias of Var. Estimators, in %, Continuous Response, SIC, $n = 720$

	25% missing					20% missing					10% missing					5% missing				
	−17					−8					0					7				
	16	−16				7	−8				0	−3				−5	2			
V1	−14	15	−15			−6	8	−10			0	3	−5			4	−1	−1		
	−2	1	−1	3		0	−2	1	8		2	−1	−1	3		2	−1	−0	−0	
	4	−2	2	−2	−20	1	1	−0	1	−17	−3	4	−4	−2	−6	−3	3	−3	0	−2
	−16					−7					0					7				
	15	−15				6	−7				0	−3				−5	2			
V3	−13	14	−13			−5	7	−9			0	3	−4			4	−1	−0		
	−2	1	−0	5		0	−2	1	9		2	−1	−1	3		2	−1	0	0	
	5	−2	2	−2	−25	2	1	−0	2	−20	−3	4	−4	−2	−7	−3	3	−3	0	−2
	−22					−12					−3					5				
	21	−22				12	−14				3	−6				−3	−0			
V4	−20	21	−21			−11	14	−15			−3	7	−8			2	1	−3		
	−0	1	−0	−7		2	−2	1	−1		3	−1	−1	−3		2	−1	−0	−4	
	1	−2	2	−3	−4	−2	1	−0	1	−3	−5	4	−4	−2	2	−4	3	−3	−0	2
	−25					−14					−3					5				
	24	−25				14	−16				4	−7				−3	−0			
V4A	−22	24	−24			−12	16	−17			−3	7	−8			2	1	−3		
	1	1	−0	−12		2	−2	1	−5		3	−1	−1	−4		2	−1	−0	−4	
	2	−2	2	−3	−10	−1	1	−0	1	−7	−5	4	−4	−2	1	−4	3	−3	−0	2
	−22					−12					−2					5				
	21	−22				12	−14				3	−6				−3	0			
V4B	−19	21	−21			−10	14	−15			−3	7	−8			2	1	−2		
	−0	1	−0	−7		2	−2	1	−1		2	−1	−1	−3		2	−1	−0	−4	
	1	−2	2	−3	−4	−2	1	−0	1	−3	−5	4	−4	−2	3	−4	3	−3	−0	2
	−25					−14					−3					5				
	24	−25				14	−15				4	−6				−3	−0			
V5	−22	24	−24			−12	16	−17			−3	7	−8			2	1	−3		
	1	1	−0	−12		2	−2	1	−5		3	−1	−1	−4		2	−1	−0	−4	
	2	−2	2	−3	−9	−1	1	−0	1	−7	−5	4	−4	−2	2	−4	3	−3	−0	2
	−24					−14					−3					5				
	22	−23				13	−14				4	−6				−3	0			
V6	−21	22	−22			−11	15	−16			−3	7	−8			2	1	−2		
	1	0	−0	−10		2	−2	1	−3		3	−1	−1	−3		2	−1	0	−3	
	3	−2	2	−2	−14	−0	1	−0	1	−10	−4	4	−4	−2	1	−4	3	−3	0	2

Table 5.13 summarizes the simulation results on the performance of variance estimator V4 (or V4B) when the response variable is continuous. The three sample sizes, 240, 720 and 1,200, and the three missing probabilities, 0.05, 0.15 and 0.25, represent three scenarios of being small, medium and large.

For simple random sampling (SRS), the relative biases are all within 10% when the missing probability is 5%. When the missing probability is 25%, the relative biases can be as large as 19% for $n = 240$, but are below 10% for almost all cases when $n = 720$ and $n = 1,200$. There is one abnormal case for β_3. The variance estimator for $\hat{\beta}_3$ has a large negative bias when the missing probability is 25%. A possible reason for this is that the true value of the variance in this case is very small, and therefore the measure of relative bias can be unreliable.

For stratified simple random sampling (STSI), the pattern of behavior of the variance estimator is similar to the cases under simple random sampling. The abnormal behaviour of the variance estimator for $\hat{\beta}_3$ is probably once again due to the small value of the true variance.

The picture for cluster sampling seems to be a little bit different, at least on the first look. The relative biases have a clear trend of decreasing as sample size increases, but the magnitude of the bias is bigger than those we see from simple random sampling and stratified sampling. However, the number of clusters under each sample size n is much smaller. For instance, when the average element sample size is 720, the actual number of clusters sampled is 108. The precision of the variance estimation depends largely on the number of clusters sampled.

There is another observation on the variance estimation under all three sampling designs. That is, the relative bias is generally bigger when the probability of missing is larger, which is a sign that the proposed variance estimators do not catch well the variance component due to imputation. This problem disappears for the alternative variance estimators presented in the next section.

Results on variance estimation under a binary response are summarized in Table 5.14. A seemingly striking pattern is that the relative biases are all within 10% when the missing probability is 5% or 15% (regardless of the sample size), and when the sample size is 1,200 (regardless of the missing probability). Overall, the relative biases are tolerable when $n \geq 720$, regardless of the sampling designs. For $n = 1,200$, the relative biases are small even if the missing probability is 25%. The performance hence is more satisfactory compared to cases with a continuous response variable.

Table 5.13: Rel. Bias of V4 (or V4B), in %, Continuous Response

		SRS (V4)					STSI (V4)					SIC (V4B)				
n	p_m	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4
		-17					-12					-24				
		14	-14				11	-13				23	-25			
	0.25	-11	13	-12			-11	13	-13			-20	23	-23		
		4	0	-2	-19		9	-5	5	-19		0	3	-3	-14	
		6	-2	0	1	-12	2	2	-2	-1	-11	-3	4	-5	-7	-8
		-12					-11					-18				
		8	-7				10	-11				17	-18			
240	0.15	-8	6	-6			-10	12	-12			-16	18	-19		
		4	-4	4	-8		10	-7	7	-19		-3	4	-4	-17	
		8	-5	4	6	-10	0	1	-2	2	-3	3	-2	3	6	-5
		-8					-4					-10				
		5	-4				3	-4				9	-9			
	0.05	-5	3	-3			-2	5	-5			-8	9	-10		
		3	-3	3	-9		10	-8	8	-17		-3	1	0	-9	
		6	-3	3	8	-10	-2	3	-3	4	-1	3	-2	2	7	-3
		-7					-8					-22				
		6	-4				7	-7				21	-22			
	0.25	-5	4	-5			-6	6	-6			-19	21	-21		
		5	-4	4	-14		-1	4	-4	-11		0	1	0	-7	
		-4	4	-4	3	0	1	1	-1	-6	-7	1	-2	2	-3	-4
		-3					-1					-11				
		2	0				1	-3				11	-14			
720	0.15	-2	0	-1			-3	5	-6			-10	14	-16		
		-1	0	0	1		1	-1	1	0		2	-1	0	-2	
		7	-7	7	-3	1	0	3	-4	0	-4	-3	2	-3	-2	0
		0					5					5				
		0	2				-4	2				-3	0			
	0.05	0	-1	1			2	0	-1			2	1	-2		
		-2	3	-3	-1		-4	3	-3	4		2	-1	0	-4	
		5	-7	7	-4	4	0	3	-3	2	-3	-4	3	-3	0	2
		-9					-7					-14				
		8	-9				4	-4				14	-16			
	0.25	-5	8	-8			-4	4	-4			-13	15	-16		
		5	-2	1	-16		5	-2	2	-7		-4	3	-3	-2	
		1	-1	0	-3	6	7	-4	3	-3	-8	1	-2	2	3	-2
		4					-7					-4				
		-4	4				4	-4				5	-7			
1200	0.15	5	-4	3			-3	3	-3			-6	7	-8		
		0	-3	4	-1		4	-1	1	-11		3	-4	4	1	
		0	0	-1	4	2	0	-1	1	0	-1	-2	2	-2	-3	2
		-2					-3					9				
		0	-1				0	1				-7	5			
	0.05	1	0	0			1	-1	0			5	-4	3		
		0	-1	3	-1		5	-1	1	-8		1	-2	3	10	
		-1	2	-3	1	1	3	-3	3	-4	1	0	-1	0	-4	2

Table 5.14: Rel. Bias of V4 (or V4B), in %, Binary Response

		SRS (V4)					STSI (V4)					SIC (V4B)				
n	p_m	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4
		-7					-16					-20				
		8	-10				14	-14				18	-17			
	0.25	-8	10	-11			-13	13	-13			-16	15	-15		
		-2	4	-4	-11		-1	5	-4	-18		3	-2	3	-9	
		4	-2	2	0	-8	11	-9	9	-2	-13	5	-5	4	-1	0
		0					-7					-2				
		1	-2				6	-6				2	-1			
240	0.15	-1	3	-4			-5	5	-4			-3	2	-3		
		-2	3	-3	-4		1	1	-1	-9		2	-2	4	-3	
		-1	4	-5	-3	-4	-1	2	-2	-4	-3	1	0	0	-1	-11
		2					0					-2				
		-2	1				0	-1				1	-1			
	0.05	1	0	0			0	1	-1			-3	2	-3		
		-3	2	-1	-1		0	4	-4	-6		4	-4	6	-2	
		-1	3	-3	-3	-4	-4	5	-5	-7	-2	5	-2	3	-5	-10
		-10					-11					-11				
		10	-11				11	-11				10	-10			
	0.25	-10	11	-11			-11	11	-10			-9	9	-8		
		-6	6	-5	-3		4	-4	4	-6		0	2	-2	-7	
		5	-3	2	-1	-9	-1	4	-4	-2	-13	1	-1	0	-1	-5
		-1					-10					-2				
		2	-4				8	-7				2	-1			
720	0.15	-2	4	-4			-6	6	-5			-1	0	1		
		-3	3	-3	-1		0	0	1	-2		4	-5	4	5	
		-5	7	-6	-1	-7	5	-5	4	2	-5	-5	4	-4	-1	2
		2					-7					4				
		0	-1				5	-4				-4	6			
	0.05	-1	2	-2			-4	3	-2			4	-6	6		
		-6	4	-4	7		0	0	0	1	-1	4	-6	5	4	
		-4	5	-5	4	-6	5	-5	5	-1	-3	-3	2	-2	0	4
		-3					-1					-2				
		3	-4				0	0				3	-6			
	0.25	-3	4	-3			1	0	-1			-4	7	-9		
		0	3	-3	-7		-2	3	-4	-2		3	0	-1	-9	
		1	-1	1	0	-4	2	-1	1	-1	-1	0	3	-3	-2	-4
		-9					3					7				
		9	-8				-5	5				-7	8			
1200	0.15	-8	7	-6			5	-4	2			6	-7	7		
		-2	1	-1	-5		-3	4	-4	0		3	-3	2	-9	
		-2	2	-1	1	3	5	-4	2	0	-3	0	1	-1	1	1
		-8					7					6				
		9	-9				-8	8				-4	3			
	0.05	-9	9	-8			8	-8	6			2	-2	1		
		-4	4	-3	-6		-4	4	-4	0		-1	0	-1	-4	
		0	0	1	-2	6	3	-3	1	3	-1	-1	1	0	2	2

5.3.3 Variance Estimation Using the Alternative Variance Decomposition

We can think of several additional ways of evaluating the variance estimator. This is so because, in section 4.4, we decomposed the variance of $\hat{\boldsymbol{\beta}}$, and its estimation, into two pieces, sampling and imputation components. The first and easiest option is to use (for the "meat") just the naive estimator, $\hat{V}_{\tau\text{naïve}}$, of the sampling component. We call this estimator AV1:

$$\text{AV1} = [\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}\hat{V}_{\tau\text{naïve}}\,[\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}.$$

Another way of estimating the total variance could be to include just an estimator of the sampling component, but correcting for the fact that some of the values in the Horvitz-Thompson estimator used are not actual observed values, but sample-based quantities. We call this estimator AV2:

$$\text{AV2} = [\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}\left(\frac{n}{r}\right)^2\hat{V}_{\tau\text{naïve}}\,[\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}.$$

A third method is to add up an estimator for each of the two pieces, sampling and imputation components, but without the correction of the sampling piece that AV2 includes. This estimator is:

$$\begin{aligned}\text{AV3} &= [\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}\left[\hat{V}_{\tau\text{naïve}} + \hat{\text{Var}}_I[U_n^*(\boldsymbol{\beta})]\right][\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}\\ &= [\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}\left[\hat{V}_{\tau\text{naïve}} + \sum_s w_i^2\frac{\partial\hat{\boldsymbol{\mu}}_i'}{\partial\hat{\boldsymbol{\beta}}}\hat{V}_i^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \text{diag}(s_{\tau r}^2)\end{pmatrix}\hat{V}_i^{-1}\frac{\partial\hat{\boldsymbol{\mu}}_i}{\partial\hat{\boldsymbol{\beta}}}\right][\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}.\end{aligned}$$

And the most complete estimator is the one that adds an estimator for each piece, but also corrects for the naïve sampling one:

$$\begin{aligned}\text{AV4} &= [\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}\left[\left(\frac{n}{r}\right)^2\hat{V}_{\tau\text{naïve}} + \hat{\text{Var}}_I[U_n^*(\boldsymbol{\beta})]\right][\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}\\ &= [\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}\left[\left(\frac{n}{r}\right)^2\hat{V}_{\tau\text{naïve}} + \sum_s w_i^2\frac{\partial\hat{\boldsymbol{\mu}}_i'}{\partial\hat{\boldsymbol{\beta}}}\hat{V}_i^{-1}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \text{diag}(s_{\tau r}^2)\end{pmatrix}\hat{V}_i^{-1}\frac{\partial\hat{\boldsymbol{\mu}}_i}{\partial\hat{\boldsymbol{\beta}}}\right][\hat{H}(\hat{\boldsymbol{\beta}})]^{-1}.\end{aligned}$$

The performance of these four alternative variance estimators is evaluated by their relative biases, calculated as for the previous nine variance estimators.

The relative bias of the four alternative estimators, AV1, AV2, AV3, and AV4, is shown in Table 5.15 for binary response, SRS sampling, and smallest sample size of $n = 240$; in Table 5.16 for binary response, STSI sampling, and largest sample size of $n = 1,200$; and in Table 5.17 for continuous response, SIC sampling, and medium sample size of $n = 720$ (other results are omitted, but the conclusions are similar).

These tables show that the estimators AV1 and AV3 perform consistently worse than the other two. We remind the reader that these estimators do not correct the naïve estimator of the sampling component; and additionally, AV1 does not include a term for the imputation component. This is an indication that *it is necessary* to "inflate" the term $\hat{V}_{\tau\text{naïve}}$ by $(n/r)^2$ in order to obtain a less biased estimator of the sampling component. Recall that not doing

so is tantamount to treating the terms $\bar{y}_{\tau r}$ in the Horvitz-Thompson estimator 4.26 as actual observations.

Additionally, AV3 performs slightly better than AV1, which indicates that, at least if one does not correct $\hat{V}_{\tau\text{naïve}}$, then it is better to include the imputation variance term.

The best estimators are AV2 and AV4, with AV4 being better than AV2. In most cases, especially for small sample sizes and/or large missing percentages, AV4 is *much better*. This means that it is generally better to *include* an estimator for the imputation component. Nonetheless, there is one situation in which it is immaterial to include it; or even better not to. It is the case where the sample size is large *and* the missing percentage is small. Even in the few situations in which AV2 is better than AV4, the relative bias of AV4 is only slightly bigger than that of AV2. And also, in those cases, AV4 tends to *overestimate*, whereas AV2 tend to *underestimate*.

In summary, AV4 is in most cases better than AV2. When the opposite happens, they are close and AV4 has a tendency to overestimate and AV2 to underestimate. Since slight overestimation of variance is preferable to slight underestimation, and the computation of the imputation variance term is not burdensome, we recommend using AV4 always.

Another advantage of AV4 over AV2 is that its relative bias does not seem to be consistently influenced by the missing percentage. As mentioned before, this is a good characteristic, and also, obviously it relies on the MAR assumption and the missing fraction being within the limits examined here.

Table 5.15: Rel. Bias of Alternative Var. Estimators, in %, Binary Response, SRS, $n = 240$

p_m	AV1					AV2					AV3					AV4				
.25	−51					−12					−35					3				
	48	−52				13	−15				34	−37				−2	1			
	−45	51	−52			−12	15	−15			−32	36	−37			1	−0	0		
	4	5	−5	−51		−1	4	−4	−13		2	5	−5	−36		−4	4	−4	2	
	13	−3	3	0	−49	4	−2	2	0	−10	10	−3	2	0	−39	1	−2	2	0	0
.20	−38					−4					−24					11				
	35	−39				3	−5				21	−24				−11	10			
	−32	38	−40			−3	5	−6			−19	24	−25			10	−9	9		
	4	2	−3	−42		−0	1	−2	−10		2	2	−2	−29		−2	1	−1	4	
	10	1	−3	−3	−41	2	2	−3	−3	−8	8	1	−3	−3	−32	0	2	−3	−3	1
.15	−33					−7					−21					5				
	31	−34				7	−9				20	−22				−4	3			
	−29	34	−35			−7	9	−10			−18	22	−23			4	−2	2		
	2	3	−4	−33		−1	2	−3	−8		0	3	−3	−22		−3	2	−3	4	
	6	3	−4	−3	−34	0	3	−4	−3	−8	4	3	−4	−3	−26	−2	4	−5	−3	−1
.10	−24					−6					−16					2				
	23	−26				7	−8				15	−17				−1	0			
	−22	26	−27			−7	9	−10			−14	17	−18			1	0	−1		
	1	3	−3	−25		−2	3	−3	−8		−0	3	−3	−17		−3	2	−2	1	
	5	2	−3	−2	−25	0	2	−3	−2	−8	3	2	−3	−2	−20	−1	3	−3	−2	−3
.05	−10					0					−6					4				
	9	−11				0	−1				5	−7				−4	3			
	−9	11	−11			−0	1	−2			−5	7	−7			3	−3	2		
	−1	1	−1	−10		−2	1	−0	−0		−1	1	−0	−6		−2	1	−0	3	
	2	2	−3	−3	−13	−1	3	−3	−3	−4	1	2	−3	−3	−11	−1	3	−3	−3	−2

Table 5.16: Rel. Bias of Alternative Var. Estimators, in %, Binary Response, STSI, $n = 1,200$

p_m	AV1					AV2					AV3					AV4				
	−50					−11					−33					6				
	46	−50				10	−11				30	−32				−7	7			
.25	−43	49	−50			−9	11	−11			−27	31	−32			7	−7	7		
	6	2	−3	−51		1	2	−2	−13		4	2	−3	−33		−2	2	−2	5	
	11	−2	1	−0	−48	3	−1	1	−0	−8	8	−1	1	−0	−37	0	−0	0	−0	3
	−42					−9					−27					6				
	39	−42				8	−10				24	−27				−6	5			
.20	−36	42	−44			−8	11	−13			−23	27	−29			5	−4	2		
	1	5	−5	−45		−3	4	−4	−14		−1	5	−5	−30		−5	4	−4	1	
	10	−3	3	−0	−41	3	−2	2	−0	−7	8	−3	2	−0	−30	1	−2	1	−0	3
	−33					−7					−20					5				
	29	−32				5	−6				17	−19				−7	7			
.15	−27	32	−34			−4	6	−8			−15	19	−21			7	−6	5		
	1	4	−4	−34		−2	3	−4	−9		−0	4	−4	−22		−4	3	−4	4	
	11	−4	3	−0	−34	5	−3	2	−0	−9	9	−4	2	−0	−26	4	−3	2	−0	−1
	−22					−3					−12					6				
	19	−21				2	−2				10	−11				−7	8			
.10	−17	21	−22			−1	3	−4			−9	11	−13			8	−7	5		
	−1	5	−5	−22		−3	5	−5	−4		−2	5	−5	−13		−4	5	−5	5	
	9	−4	2	1	−25	5	−3	2	1	−7	8	−3	2	1	−19	4	−3	1	1	−2
	−8					2					−3					7				
	6	−7				−3	3				1	−1				−8	9			
.05	−5	7	−9			4	−2	1			−0	2	−3			8	−8	6		
	−1	4	−4	−14		−3	4	−4	−4		−2	4	−4	−9		−3	4	−4	1	
	6	−3	2	3	−13	4	−3	2	3	−4	5	−3	2	3	−11	3	−3	2	3	−1

Table 5.17: Rel. Bias of Alternative Var. Estimators, in %, Continuous Response, SIC, $n = 720$

p_m	AV1					AV2					AV3					AV4				
	−52					−14					−41					−3				
	49	−53				15	−16				38	−41				4	−4			
.25	−45	51	−52			−14	15	−15			−35	39	−40			−4	3	−3		
	7	−0	0	−53		3	−0	0	−17		5	−0	0	−35		−0	0	0	2	
	9	−2	2	−2	−49	2	−2	2	−2	−9	7	−2	2	−2	−39	0	−2	2	−2	1
	−39					−5					−29					5				
	37	−41				6	−8				27	−30				−4	3			
.20	−33	41	−43			−6	9	−10			−24	30	−32			3	−2	1		
	8	−3	1	−44		4	−3	2	−12		6	−2	1	−27		1	−2	1	5	
	5	1	−1	0	−41	−1	1	−0	0	−8	4	1	−1	1	−33	−3	2	−1	1	0
	−31					−5					−23					3				
	30	−34				6	−9				22	−26				−2	−1			
.15	−27	35	−36			−7	10	−12			−20	26	−28			0	2	−3		
	7	−2	0	−36		4	−2	1	−12		5	−1	0	−22		2	−1	1	2	
	3	2	−2	−2	−31	−1	2	−2	−2	−4	2	2	−2	−2	−24	−3	2	−2	−2	2
	−18					1					−12					7				
	18	−22				0	−3				12	−15				−5	3			
.10	−16	22	−24			−1	5	−6			−11	16	−17			4	−2	0		
	7	−1	−1	−28		4	−1	−0	−11		5	−1	−1	−18		3	−1	−1	−1	
	−1	4	−4	−2	−20	−4	4	−4	−2	−1	−2	4	−4	−2	−15	−5	4	−4	−2	3
	−4					7					−0					10				
	5	−9				−4	1				2	−5				−7	5			
.05	−5	10	−11			3	0	−2			−2	6	−8			6	−3	2		
	5	−2	0	−18		4	−2	0	−9		4	−1	0	−12		3	−1	0	−3	
	−2	3	−3	−0	−10	−4	3	−3	−0	0	−2	3	−3	−0	−7	−4	3	−3	−0	3

In Table 5.18 and Table 5.19 we present the relative bias of AV4 for both, continuous or binary response; for SRS, STSI, or SIC; for the smallest (240) medium (720) and largest (1,200) sample sizes; and for the largest (0.25), medium (0.15), and smallest (0.05) missing probabilities.

We confirm that, conditional on a given sample size, the performance of this variance estimator does not seem to be influenced by the missing probability.

For the continuous response with SRS and the binary response with STSI or SIC, the largest negative bias is -9% and the largest positive is about 13%. For the continuous response with STSI, the largest negative is -13% and the largest positive is 7%. For the continuous response with SIC, the largest negative is -8% and the largest positive 12%. For the binary response with SRS, the largest negative is -10% and the largest positive is 11%. And for the continuous response with SIC, the largest negative and positive relative biases are -12% and 14%, respectively.

It is clear that in all instances considered, this alternative estimator performs (much) better than the estimator in the previous section. Obviously this variance estimator does a much better job at handling both the variability due to the imputation and the sampling variability.

This variance estimator does not seem to have a consistently negative or positive bias. Over all the scenarios and simulations considered, the highest and lowest relative biases were 14% and -13%, respectively. And in may cases, for a particular response variable, sampling design, sample size, and missing fraction, the relative bias of the variance estimator is positive for some β's and negative for others.

Although, in general, there appears to be a decreasing trend in the (absolute value of the) relative bias of AV4 as the sample size increases, this tendency is not monotone. In several cases the maximum relative bias (in absolute value) increases for a larger sample size. At the moment we do not have a satisfactory explanation for this fact. It may be that the rate at which the bias of AV4 decreases is not the same as the rate at which the actual MSE does. Since the relative bias has the MSE as denominator, this could explain this phenomenon.

We observe that in some scenarios, here the variance estimator corresponding to $\hat{\beta}_3$ also shows an irregular behaviour (compared to the others'). This further supports the hypothesis that since the true MSE of $\hat{\beta}_3$ is close to zero, the measure of relative bias used may not be as stable for this estimator as it is for the others.

Table 5.18: Rel. Bias of Alternative Var. Estimator AV4, in %, Continuous Response

		SRS					STSI					SIC				
n	p_m	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4
		-5					6					-2				
		2	-1				-5	5				3	-5			
	0.25	-1	0	1			4	-4	4			-2	4	-3		
		0	2	-3	-2		4	-4	4	1		-1	3	-3	-2	
		5	-2	1	1	-6	1	2	-2	0	-4	-4	4	-4	-7	-2
		-7					-4					-6				
		4	-1				4	-4				5	-6			
240	0.15	-4	1	-1			-4	5	-6			-6	7	-7		
		2	-4	4	-2		8	-7	7	-10		-3	3	-4	-12	
		8	-5	4	6	-8	-1	1	-2	2	0	2	-2	3	6	-3
		-5					2					-4				
		3	-1				-2	1				3	-4			
	0.05	-2	0	0			2	0	-1			-4	4	-4		
		2	-3	3	-6		9	-8	7	-12		-3	1	0	-5	
		5	-3	3	8	-8	-3	3	-3	5	2	2	-1	2	7	-2
		3					1					-3				
		-4	6				-2	3				4	-4			
	0.25	3	-5	4			2	-3	4			-4	3	-3		
		3	-3	3	-5		-3	4	-4	0		0	0	0	2	
		-5	4	-4	3	4	-1	2	-2	-6	-2	0	-2	2	-2	1
		1					3					3				
		-1	4				-2	1				-2	-1			
720	0.15	1	-3	3			0	1	-2			0	2	-3		
		-1	-1	1	4		-1	0	1	5		2	-1	1	2	
		6	-8	8	-3	3	-1	3	-3	1	-2	-3	2	-2	-2	2
		1					4					10				
		-1	3				-3	2				-7	5			
	0.05	0	-2	2			2	0	-1			6	-3	2		
		-2	2	-3	0		-4	3	-3	5		3	-1	0	-3	
		5	-7	7	-4	4	0	3	-4	2	-3	-4	3	-3	0	3
		-2					2					7				
		1	-1				-4	5				-5	5			
	0.25	1	0	1			5	-5	5			3	-4	3		
		3	-1	2	-8		2	-1	1	2		-5	3	-2	8	
		0	-1	0	-3	10	6	-4	3	-3	-4	1	-2	2	3	3
		7					-3					11				
		-8	7				1	0				-8	7			
1200	0.15	8	-7	6			-1	0	0			6	-6	5		
		-2	-2	4	3		3	-1	1	-7		4	-4	4	6	
		0	0	-1	4	4	0	-1	1	0	1	-2	2	-1	-4	3
		-1					-2					14				
		-1	1				-1	2				-11	10			
	0.05	2	-1	1			1	-1	0			9	-9	8		
		-1	-1	3	0		5	-1	1	-7		1	-2	3	12	
		-1	2	-3	2	1	3	-3	3	-4	1	0	-1	1	-4	3

Table 5.19: Rel. Bias of Alternative Var. Estimator AV4, in %, Binary Response

		SRS					STSI					SIC				
n	p_m	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4	β_0	β_1	β_2	β_3	β_4
		3					-3					-8				
		-2	1				2	-1				7	-5			
	0.25	1	0	0			-2	1	0			-5	4	-3		
		-4	4	-4	2		-3	4	-3	-4		1	-2	3	4	
		1	-2	2	0	0	9	-9	8	-1	-3	3	-5	4	-1	9
		5					-2					5				
		-4	3				1	0				-5	5			
240	0.15	4	-2	2			0	0	1			3	-4	3		
		-3	2	-3	4		0	1	-1	-2		0	-1	3	3	
		-2	4	-5	-3	-1	-1	2	-2	-5	-1	1	0	1	-2	-8
		4					3					2				
		-4	3				-3	2				-2	3			
	0.05	3	-3	2			3	-2	3			0	-2	1		
		-2	1	0	3		-1	4	-4	-1		3	-3	5	1	
		-1	3	-3	-3	-2	-5	5	-5	-7	2	4	-2	3	-5	-8
		-1					-2					-1				
		2	-3				2	-2				1	1			
	0.25	-3	3	-3			-2	2	-1			0	-2	3		
		-7	6	-5	5		2	-3	3	2		-2	2	-2	2	
		3	-2	2	-1	-4	-2	4	-3	-2	-7	0	-1	0	-2	0
		3					-7					4				
		-2	0				4	-3				-4	4			
720	0.15	2	0	1			-3	2	-1			4	-5	6		
		-4	3	-3	2		0	-1	1	1		2	-4	3	10	
		-6	8	-7	0	-4	4	-4	4	2	-3	-5	4	-4	-2	4
		3					-6					6				
		-1	-1				4	-3				-6	7			
	0.05	0	1	-1			-3	2	-2			6	-7	8		
		-6	5	-4	8		0	0	0	2		4	-5	4	6	
		-4	5	-5	4	-5	5	-5	5	-1	-2	-3	2	-2	-1	4
		3					6					4				
		-2	2				-7	7				-3	1			
	0.25	3	-3	4			7	-7	7			3	-1	1		
		0	2	-2	0		-2	2	-2	5		1	1	-3	6	
		0	0	0	-1	1	0	0	0	0	3	-1	3	-4	-3	0
		-7					5					4				
		7	-7				-7	7				-1	-1			
1200	0.15	-6	5	-4			7	-6	5			0	2	-2		
		-2	1	0	-2		-4	3	-4	4		-3	3	-3	6	
		-2	1	-1	1	6	4	-3	2	0	-1	-4	4	-5	-3	7
		-8					7					7				
		9	-9				-8	9				-4	0			
	0.05	-9	8	-8			8	-8	6			3	0	0		
		-4	3	-2	-6		-3	4	-4	1		-4	1	0	11	
		0	0	0	-2	6	3	-3	2	3	-1	-9	10	-10	-1	4

Chapter 6

Conclusions and Future Research

6.1 Concluding Remarks

In this thesis we argue that the most appropriate way of inference for analytical purposes based on complex survey data is a joint randomization approach; where all sources of randomness are considered together in order to draw inferences.

Under such perspective we showed that the pseudo-GEE with complete survey data provides consistent estimators of superpopulation parameters; and presented the break-down of the variance of the estimator, into a sampling and an model variance component.

Using the same framework, we also showed how the weighted GEE proposed by Robins et al. (1995), to deal with dropouts and missing waves in longitudinal studies, can be extended to handle the complex survey situation. It turns out, in this case the method uses what we called "wave-specific" survey weights in the estimating equation, to get the point estimators.

We proposed a hot-deck imputation scheme, and a corresponding hot-deck imputation pseudo-GEE, to deal with missing *responses* in longitudinal surveys, when there is no missing values in the covariates, and all covariates are categorical. We showed that the estimators obtained by such a method are consistent for the superpopulation parameters when the missing values occur at random (MAR). The consistency of the estimator is confirmed in the simulation studies; the results also revealed that the estimator has good finite sample properties.

The simulations evaluated the performance of the proposed estimator for a continuous response and a binary response, and for different sampling designs. We found that, in all cases, the bias of the estimator is pretty small even for the small sample sizes and the largest missing fractions considered.

Additionally, we developed mathematical expressions for the variance, and their corresponding estimators, for the proposed point estimator, in two different ways. The first one decomposed the total variance into three pieces, corresponding to the sampling variability, the imputation variability, and a mixed component.

Based on the simulations we concluded that, although the expression for the variance contains some double summation terms (over the sample), it is actually not required to calculate these terms for the element sampling cases considered. On the other hand, for the cluster sampling case it is necessary to compute a "between cluster" double summation over pairs of

elements belonging to different clusters. But still, it is not necessary to compute any summations for pairs of elements belonging to the same cluster.

The simulation results also showed that this variance estimator performed well for the binary response, but not so well for the continuous response. And the variance estimator has the good property that its bias has a tendency to reduce as the sample size increases. However, it has the drawback that in many cases, the larger the missing fraction, the larger the bias. We conclude that this variance estimator does not capture the variability due to imputation quite well.

The second variance decomposition and estimator, breaks up the total variability into a sampling component and an imputation component. It makes an ad-hoc correction to a simple variance estimator, does not require any double summation terms, and is easier and faster to compute.

The simulation results show that this variance estimator performs generally pretty well, and better than the first variance estimator. It has the nice feature that its bias does not depend on the missing percentage. And the bias of this variance estimator does exhibit a decreasing trend as the sample size increases, although this trend is not as pronounced as it for the previous variance estimator.

6.2 Future Research

There are several directions that this thesis work can be extended or continued. The simulation results reported in chapter 5 used the unweighted hot-deck imputation method. Repeating the study with weighted hot-deck imputation procedure would be of some interest. In addition, examination on the performance of the proposed estimation strategies under sampling with selection probability proportional to a size measure would also be of interest. There are several other less trivial aspects and/or directions I would like to pursue in future research.

(i) Simulation studies on the performance of the pseudo GEE estimators under imputation for other types of response variables. Count or ordinal response variables will be of primary interest.

(ii) Comparisons between the reweighing method of Robins et al. (1995) and the imputation method proposed in this thesis.

(iii) Hot-deck imputation when some of the covariates used in the model are continuous. An ad-hoc procedure in this case is to categorize the involved continuous covariates when formulating imputation classes, and then use weighted or unweighted hot-deck method to impute missing responses. It will have both practical and theoretical values in exploring the impact of such a discretization of covariates on the estimation of model parameters as well as on variance estimation.

A more attractive approach in the presence of continuous covariates is perhaps to use the nearest neighbour imputation method. There are several related issues here. The very first one is how much gain in efficiency from using the latter as compared to the former

which is operationally simple; the second is the consistency of the resulting pseudo GEE estimators under nearest neighbour imputation; and the third is the variance estimation under the latter approach.

(iv) Theoretical justification of the alternative variance estimation method presented in section 4.4. The ad-hoc adjustment by $(n/r)^2$ seems to work well in the simulation studies reported in chapter 5. This method is considerably simpler and is easy to implement, but it requires further confirmation from theory before the method can be recommended for longitudinal analysis.

(v) Variance estimators based on replicated bootstrap weights. There has been increased interest in recent years among survey practitioners, especially those from Statistics Canada and other large organizations, to use variance estimators based on replicated survey weights, for which bootstrap variance estimators are the most popular ones. To develop such variance estimation techniques for the pseudo GEE estimators under imputation for missing responses will be a challenge and yet important task for the analysis of complex longitudinal surveys.

(vi) Handling missing values in both the response variable and covariates. The difficulty of this problem, which is quite common for all types of large scale surveys, becomes much more pronounced for longitudinal surveys due to the complex structure over several time points.

Appendix A

Detailed Description of the NLSCY

SAMPLING DESIGN

Target population. The NLSCY only surveys the non-institutionalized civilian population (aged 0 to 11 at the time of their selection) in Canada's 10 provinces. The survey excludes children living on Indian reserves or Crown lands, residents of institutions, full-time members of the Canadian Armed Forces, and residents of some remote regions. The sample is intended to represent the population for both longitudinal and for cross-sectional purposes, at the time of collection.

Sampling unit. "Most samples were drawn from the Labour Force Survey's (LFS) sample of respondent households, with the exception of one-year-olds sampled in 1998 and the five year-olds sampled in 2000 who were selected using provincial birth registry data since the LFS did not have enough eligible children to meet the survey's needs" (Statistics Canada, Cycle4). So, we have that for some cases (the ones from LFS) the sampling unit is the dwelling (not the household), and in other cases (the ones from the birth registry) it is the child. (Even though the observation unit is always the child). For our study, all the observations are from the first cohort. So, all of them are selected from the LFS and then the sampling unit is the dwelling.

Sample size (Only longitudinal components for first four cycles are given since our study concentrates only in this part). Cycle 1: Intended: 15,579 households [dwellings]; Achieved: 13,439 households [dwellings]; Achieved kids: 22,831 children 0 to 11. Cycle 2: Intended: 16,903 children (out of the 22,831; limiting to at most 2 children per hh); Achieved: 15,468 children 2 to 13. Cycle 3: Intended: 16,903 children Achieved: 14,997 children. Cycle 4: Intended: 16,903 children; Achieved: 13,310 children. Respondents to all cycles; Achieved: 11,136 children

Collection period. Cycle 1: from Dec 1994 to April 1995. Cycle 2: from Dec 1996 to April 1997. Cycle 3: from Oct 1998 to June 1999. Cycle 4: from Sept 2000 to May 2001.

Stratification.

Primary strata: Provinces are divided into economic regions and employment insurance economic regions. Economic Regions (ERs) are geographic areas of more or less homogeneous economic structure. Employment insurance economic regions (EIER) are also geographic areas, and are roughly the same size and number as ERs, but they do not share the same definitions. The intersections of the two types of regions form the primary strata. Census

Metropolitan Areas (CMAs) are also an EIERs. These strata are classified into three types of areas: rural, urban, and remote areas.

Secondary strata: In urban areas with sufficiently large numbers of apartment buildings, the strata are grouped according to those based on apartment frames and those based on area frames. [Within each secondary stratum further stratification is carried out to reflect differential population concentration and socio-economic characteristics]. In rural areas, stratification strategies were based not only on concentration of population, but also on cost-efficiency and interviewer constraints. [Further stratification to reflect differences among a number of socio-economic characteristics within each stratum]. The remote area frame is stratified only by province.

Allocation of sample to strata. The NLSCY sample for Cycle 1 was constructed taking two important requirements into consideration. A sufficient sample was required in each of the 10 provinces to allow for the production of reliable estimates for all children 0 to 11 years of age. The sample allocation was derived such that the smaller provinces had sufficient sample to meet this requirement.

A second requirement was that it was necessary to have a large enough sample to produce estimates at the Canada level by seven key age groupings or cohorts: 0 to 11 months, 1, 2 to 3, 4 to 5, 6 to 7, 8 to 9, and 10 to 11 years. It was possible to over sample households which contained at least one child in the youngest two age groupings to allow for the sample requirements for these age groups.

Sample clustering.

Each stratum is divided into clusters, and then a sample of clusters is selected within the stratum. Dwellings are then sampled from selected clusters. Within each urban stratum in the urban area frame, a number of geographically contiguous groups of dwellings, or clusters, are formed based upon 1991 Census counts. The selection of a sample of clusters (always six or a multiple of six clusters) from each of these secondary strata represents the first stage of sampling in most urban areas. In some other urban areas, census enumeration areas (EA) are used as clusters. In the low density urban strata, a three stage design is followed. Under this design, two towns within a stratum are sampled, and then a multiple of six clusters within each town are sampled. For urban apartment strata, instead of defining clusters, the apartment building is the primary sampling unit. Apartment buildings are sampled from the list frame with probability proportional to the number of units in each building. Within each rural (or remote area) stratum, six EAs or two or three groups of EAs are sampled as clusters, whereas remote settlements within each province are sampled proportional to the number of dwellings in the settlement. In all three types of areas (urban, rural and remote areas) a listing of all private dwellings in the cluster is prepared. From the listing, a sample of dwellings is then selected. The sample yield depends on the type of stratum. For example, in the urban area frame, sample yields are either six or eight dwellings, depending on the size of the city. In the urban apartment frame, each cluster yields five dwellings, while in the rural areas and urban EAs, each cluster yields 10 dwellings. In all clusters, dwellings are sampled systematically. This represents the final stage of sampling [of households].

Selection of individuals. For Cycle 1, up to 4 eligible children were selected at random in each selected household; however, for Cycle 2 from cycle two onwards only two children were

selected in a household with more than two children interviewed in Cycle 1. "The longitudinal sample will be comprised of all children sampled for Cycle 1 of the survey in responding households. The plan is to follow these children over time every two years" (Statistics Canada, cycle 1).

Breakdown of responding children by province at Cycle 1.

Newfoundland: 1,232
Prince Edward Island: 764
Nova Scotia: 1,532
New Brunswick: 1,426
Québec: 4,065
Ontario: 6,020
Manitoba: 1,789
Saskatchewan: 1,878
Alberta: 2,185
British Columbia: 1,940
TOTAL: 22,831

Approximately self-weighting at household level within strata. For the urban strata, each cluster has approximately the same number of dwellings ("are formed based upon 1991 Census counts") and these clusters are selected by SRS (constant probability for each cluster). Then, 6 or 8 dwellings are selected, presumably by SRS, which is constant again; so, we are left with (approximately) constant inclusion probability for every dwelling unit. The same argument applies to urban EAs and rural areas, assuming the EAs from the census are roughly of the same size. In the low density urban strata it is not very clear but if we assume that the number of clusters selected in a town (6, 12, 18,...) is approximately proportional to the number of cluster, then this can also be thought of as approximately self-weighting. In the urban apartment strata, a big building is more probable to be selected; but since 5 units are always selected within, these units have less chance of being selected if they are in big buildings, and the two steps compensate each other.

CONSTRUCTION OF SURVEY WEIGHTS

Basic weights. Reciprocals of selection probabilities adjusted for household non-response (at cycle 1). The non-response adjustment used weighting adjustment cells defined using the following information: province, economic region, census metropolitan area, type of sector (urban, rural), apartment frame, whether special region or not. Each of the strata had at least 30 children and a response rate of at least 70%. Strata that were too small or had a response rate of less than 30% were grouped until these restrictions were met.

Extra correction. Correction for households with more than two eligible children. Done from cycle 1 to cycle 2.

Correction for attrition. At each cycle (2, 3, 4), the previous cycle's weight is adjusted for attrition within cells in that cycle, which are formed using all the available information from the earlier cycles. For example, at cycle 3, adjustment classes are formed using information from cycles 1 and 2, and in each cell, cycle 2's weight is adjusted to compensate for the attrition in cycle 3 in that cell.

Benchmarking or calibration. After the correction for attrition at cycle 4 has take place, this last adjustment factor ensures consistency between the estimates produced by the survey and Statistics Canada's population estimates (poststratification). The target population is the set of all children between the ages of 0 and 11 at the beginning of 1995. The poststratification adjustment of that sample ensures consistency between the sum of the weights and the January 1995 population estimate for each province-age-sex combination.

NON-SAMPLING ERRORS AND DATA QUALITY

Mode of data collection. Households in which all the selected children were aged 3 or under: The computer-assisted interview and the paper questionnaire on Ages and Stages were completed by telephone since neither the child nor the parent's consent and signature were required for questionnaire administration. The interview was conducted in two stages. During the initial call, the interviewer completed the computer-assisted interview and determined which version of the Ages and Stages questionnaire should be used. The interviewer told the respondent that a questionnaire would be mailed to him/her, and made an appointment to call one or two weeks later to collect the responses.

Households in which the selected children were aged 4 or over: The first few components of the computer-assisted interview were completed by telephone; the rest of the interview, which had both computer-assisted and paper components, had to be completed during a field visit. Between the initial call and the field visit, the parents of the 4-5 subgroup also received the appropriate version of the Ages and Stages questionnaire by mail so that they could complete it before the interviewer's visit.

Effort to contact, follow-up. All cases not processed in early stages of each cycle for reasons such as no contact, hard refusal or language barriers were returned to the interviewers for inclusion in a new phase sample.

Who is the respondent (proxy). In each NLSCY household, for each selected child, a question was asked about who in the household was the person most knowledgeable about this child. This person was labeled as the PMK. The PMK provides the information for all selected children in the household and then gives information about himself/herself and his/her spouse/partner.

Appendix B

NLSCY Covariates Description

- "Age": Age of the child (002-011). Treated as continuous variable.
- "Age^2": The square of the age of the child (002-011). Treated as continuous variable.
- "DeprePMK": Depression score of the person most knowledgeable about the child (PMK), ranging from 0 to 36 - a high score indicates the presence of depression symptoms. Treated as continuous variable.
- "Punitive": Punitive (aversive) parenting score (asked for children from 2-11 years old), ranging from 0 to 19 - a high score indicates punitive/aversive interactions [between parent and kid]. Treated as continuous variable.
- "Region": Region of residence (at baseline). A categorical variable with five categories: Ontario, Quebec, British Columbia, Prairie (Alberta, Manitoba, Saskatchewan), and Atlantic (New Brunswick, Newfoundland, Nova Scotia, Prince Edward Island).
This variable is entered in our models as four dummy variables, as follows:
if Ontario then Region_1=0, Region_2=0, Region_3=0, Region_4=0;
if Quebec then Region_1=0, Region_2=0, Region_3=0, Region_4=1;
if British Columbia then Region_1=0, Region_2=0, Region_3=1, Region_4=0;
if Prairie then Region_1=0, Region_2=1, Region_3=0, Region_4=0;
if Atlantic then Region_1=1, Region_2=0, Region_3=0, Region_4=0.
- "GENDER": Gender of child - Cycle 1 gender of the child: Male, Female. Categorical variable, two categories.
This variable is entered in our models as one dummy variable, as follows:
if Male then AMMCQ02_F=0;
if Female then AMMCQ02_F=1.
- "FamStat": Family Status. Child's Parent Status - Child lives with: Both biological parents, One single biological parent, One biological parent and one step parent, Other. Categorical variable with four categories.
This variable is entered in our models as three dummy variables, as follows:
if Both biological parents then FamStat_1=0, FamStat_2=0, FamStat_3=0;
if One single biological parent then FamStat_1=0, FamStat_2=0, FamStat_3=1;
if One biological & one step parent then FamStat_1=0, FamStat_2=1, FamStat_3=0;
if Other then FamStat_1=1, FamStat_2=0, FamStat_3=0.

- “Income”: Household Income Status. Ratio of income to the low-income cut-off (LICO) for the economic family: Less than 1.0, More than or equal to 1.0. Categorical variable with two categories.
This variable is entered in our models as one dummy variable, as follows:
if More than or equal to 1.0 then LowIncome_1=0;
if Less than 1.0 then LowIncome_1=1.
- “Hours”: Number of hours in daycare. Number of hours per week spent in all care arrangements. A categorical variable, with three categories: Not in daycare, Less than 50 hours, 50 hours or more.
This variable is entered in our models as two dummy variables, as follows:
if Not in child care then Hours_1=0, Hours_2=0;
if Less than 50 then Hours_1=0, Hours_2=1;
if 50 or more then Hours_1=1, Hours_2=0.
- “Age*Puni”: Age by Punitive Parenting Interaction. A continuous variable equal to the product of the two variables Age and Punitive.
- “Age*Inco”: Age by Household Income Status interaction. A continuous variable equal to the product of the two variables Age and Income.
It is equal to Age if LowIncome_1=1 and equal to zero if LowIncome_1=0.
- “Age*Regi”: Age by Region interaction. Four continuous variables equal to the product of the two variables Age and Region, as follows:
if Ontario, AgeRegion_1=0, AgeRegion_2=0, AgeRegion_3=0, AgeRegion_4=0;
if Quebec, AgeRegion_1=0, AgeRegion_2=0, AgeRegion_3=0, AgeRegion_4=Age;
if BC, AgeRegion_1=0, AgeRegion_2=0, AgeRegion_3=Age, AgeRegion_4=0;
if Prairie, AgeRegion_1=0, AgeRegion_2=Age, AgeRegion_3=0, AgeRegion_4=0;
if Atlantic, AgeRegion_1=Age, AgeRegion_2=0, AgeRegion_3=0, AgeRegion_4=0.

References

Albert, Paul S. (1999), 'Longitudinal data analysis (repeated measures) in clinical trials', *Statistics in Medicine* **18**, 1707–1732. [18]

Ali, Mirza and Enayet Talukder (2005), 'Analysis of longitudinal binary data with missing data due to dropouts', *Journal of Biopharmaceutical Statistics* **15**(6), 993–1007. [25]

Baker, Stuart G. (1995), 'Marginal regression for repeated binary data with outcome subject to non-ignorable non-response', *Biometrics* **51**(3), 1042–1052. [26]

Baker, Stuart G. and Nan M. Laird (1988), 'Regression analysis for categorical variables with outcome subject to nonignorable nonresponse', *Journal of the American Statistical Association* **83**, 62–69. [16]

Beaumont, Jean-François (2005), 'On the use of data collection process information for the treatment of unit nonresponse through weight adjustment', *Survey Methodology* **31**(2), 227–231. [16]

Binder, David A. (1983), 'On the variances of asymptotically normal estimators from complex surveys', *International Statistical Review* **51**, 279–292. [32]

Binder, David A. and Georgia R. Roberts (2003), *Analysis of Survey Data*, Wiley Series in Survey Methodology, Wiley, Chichester, chapter Design-based and Model-based Methods for Estimating Model Parameters. [27, 28]

Binder, David A. and Weimin Sun (1996), Frequency valid multiple imputation for surveys with a complex design, *in* 'ASA Proceedings of the Section on Survey Research Methods', American Statistical Association, pp. 281–286. [25]

Binder, David A. and Zdenek Patak (1994), 'Use of estimating functions for estimation from complex surveys', *Journal of the American Statistical Association* **89**(427), 1035–1043. [29]

Brick, J. Michael, Graham Kalton and Jae Kwang Kim (2004), 'Variance estimation with hot deck imputation using a model', *Survey Methodology* **30**(1), 57–66. [43, 48, 86]

Carey, Vincent, Scott L. Zeger and Peter Diggle (1993), 'Modelling multivariate binary data with alternating logistic regressions', *Biometrika* **80**(3), 517–526. [69]

Carrillo-Garcia, Ivan A. (2006), Analysis of longitudinal survey data with missing observations: An application of weighted GEE to the national longitudinal survey of children and youth (NLSCY), Technical report, Statistics Canada. MITACS/NPCDS Internship Program. [4, 62, 69]

Carrillo, Ivan, Christina Chu, Wanhua Su and Xinlei Xie (2005), A longitudinal study of factors affecting children's behaviour, Proceedings of the Survey Methods Section, Statistical Society of Canada, Saskatoon. [3, 4, 62, 69]

Carrillo, Ivan, Milorad Kovacevic and Changbao Wu (2006), Analysis of longitudinal survey data with missing observations: An application of weighted GEE to the national longitudinal survey of children and youth (NLSCY), Proceedings of the Survey Methods Section, Statistical Society of Canada, London. [62, 69]

Chen, Hua Yun (2004), 'Nonparametric and semiparametric models for missing covariates in parametric regression', *Journal of the American Statistical Association* **99**(468), 1176–1189. [16]

Chen, Jiahua and Jun Shao (2000), 'Nearest neighbor imputation for survey data', *Journal of Official Statistics* **16**(2), 113–131. [46]

Chen, Jiahua, Mary E. Thompson and Changbao Wu (2004), 'Estimation of fish abundance indices based on scientific research trawl surveys', *Biometrics* **60**(1), 116–123. [28]

Cook, Richard J., Leilei Zeng and Grace Y. Yi (2004), 'Marginal analysis of incomplete longitudinal binary data: A cautionary note on locf imputation', *Biometrics* **60**(3), 820–828. [25]

Diggle, Peter, Patrick Heagerty, Kung-Yee Liang and Scott Zeger (2002), *Analysis of Longitudinal Data*, 2 edn, Oxford University Press, New York. [1, 18, 19]

Duncan, Greg J. and Graham Kalton (1987), 'Issues of design and analysis of surveys across time', *International Statistical Review* **55**(1), 97–117. [2]

Fay, Robert E. (1996), 'Alternative paradigms for the analysis of imputed survey data', *Journal of the American Statistical Association* **91**, 490–498. [25, 26]

Fitzmaurice, Garrett M., Geert Molenberghs and Stuart R. Lipsitz (1995), 'Regression models for longitudinal binary responses with informative drop-outs', *Journal of the Royal Statistical Society, Series B: Methodological* **57**, 691–704. [16]

Fitzmaurice, Garrett M., Nan M. Laird and Gwendolyn E. P. Zahner (1996), 'Multivariate logistic models for incomplete binary responses', *Journal of the American Statistical Association* **91**(433), 99–108. [16, 26]

Fitzmaurice, Garrett M., Nan M. Laird and James H. Ware (2004), *Applied Longitudinal Analysis*, Wiley Series in Probability and Statistics, Wiley, Hoboken. [8, 10, 11, 12, 16, 18]

Galecki, Andrzej T., Thomas R. Ten Have and Geert Molenberghs (2001), 'A simple and fast alternative to the EM algorithm for incomplete categorical data and latent class models', *Computational Statistics & Data Analysis* **35**(3), 265–281. [26]

Godambe, V. P. (1995), 'Estimation of parameters in survey sampling: Optimality', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **23**(3), 227–243. [29]

Godambe, V. P. and M. E. Thompson (1986), 'Parameters of superpopulation and survey population: Their relationships and estimation', *International Statistical Review* **54**(2), 127–138. [29]

Groves, Robert M., Don A. Dillman, John L. Eltinge and Roderick J. A. Little, eds (2002), *Survey Nonresponse*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York. [2]

Hájek, Jaroslav (1960), 'Limiting distributions in simple random sampling from a finite population', *Publications of the Mathematics Institute of Hungarian Academy of Science* **5**, 361–375. [32]

Hájek, Jaroslav (1964), 'Asymptotic theory of rejective sampling with varying probabilities from a finite population', *The Annals of Mathematical Statistics* **35**, 1491–1523. [32]

Hardin, James W. and Joseph Hilbe (2003), *Generalized Estimating Equations*, Chapman & Hall Ltd. [19, 22]

Hardin, James W. and Joseph M. Hilbe (2001), *Generalized Linear Models and Extensions*, Stata Corporation, College Station. [8, 9]

Hedeker, Donald and Robert D. Gibbons (2006), *Longitudinal Data Analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken. [1]

Horvitz, D. G. and D. J. Thompson (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**(260), 663–685. [30]

Inagaki, Nobuo (1973), 'Asymptotic relations between the likelihood estimating functions and the maximum likelihood estimators', *Annals of the Institute of Statistical Mathematics* **25**, 1–26. [29]

Kalton, Graham (1983), 'Models in the practice of survey sampling', *International Statistical Review* **51**, 175–188. [28]

Kalton, Graham and Daniel Kasprzyk (1986), 'The treatment of missing survey data', *Survey Methodology* **12**, 1–16. [16]

Kim, Jae Kwang, Michael J. Brick, Wayne A. Fuller and Graham Kalton (2006), 'On the bias of the multiple-imputation variance estimator in survey sampling', *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **68**(3), 509–521. [26]

Kish, Leslie (1987), *Statistical Design for Research*, John Wiley & Sons, New York. [1, 2]

Korn, Edward L. and Barry I. Graubard (1999), *Analysis of Health Surveys*, John Wiley & Sons, New York. [2]

Kovacevic, Milorad S. and Shesh N. Rai (2002), 'Log-linear modelling of change using longitudinal survey data', *Communications in Statistics - Theory and Methods* **31**(10), 1815–1835. [31]

Lavori, Philip W., Ree Dawson and David Shera (1995), 'A multiple imputation strategy for clinical trials with truncation of patient data', *Statistics in Medicine* **14**, 1913–1925. [25]

Lehmann, Erich Leo and George Casella (1998), *Theory of Point Estimation*, 2 edn, Springer-Verlag, New York. [10]

Liang, Kung-Yee and Scott L. Zeger (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22. [11, 63]

Liang, Kung-Yee, Scott L. Zeger and Bahjat Qaqish (1992), 'Multivariate regression analyses for categorical data', *Journal of the Royal Statistical Society, Series B: Methodological* **54**(1), 3–40. [69, 72]

Lipsitz, Stuart R., Nan M. Laird and David P. Harrington (1991), 'Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association', *Biometrika* **78**(1), 153–160. [69, 72]

Little, Roderick J. A. (1992), 'Regression with missing X's: A review', *Journal of the American Statistical Association* **87**, 1227–1237. [16]

Little, Roderick J.A. and Donald B. Rubin (2002), *Statistical Analysis with Missing Data*, 2 edn, Wiley, Hoboken. [15]

Little, Roderick J.A. and Sonya Vartivarian (2005), 'Does weighting for nonresponse increase the variance of survey means?', *Survey Methodology* **31**(2), 161–168. [16]

Liu, Guanghan and A. Lawrence Gould (2002), 'Comparison of alternative strategies for analysis of longitudinal trials with dropouts', *Journal of Biopharmaceutical Statistics* **12**(2), 207–226. [25]

Lohr, Sharon L. (1999), *Sampling: Design and Analysis*, Duxbury Press, Pacific Grove. [5]

Nelder, J. A. and R. W. M. Wedderburn (1972), 'Generalized linear models', *Journal of the Royal Statistical Society A* **135**, 370–384. [8]

Parzen, Michael, Stuart R. Lipsitz, Joseph G. Ibrahim and Steven Lipshultz (2002), 'A weighted estimating equation for linear regression with missing covariate data', *Statistics in Medicine* **21**(16), 2421–2436. [16]

Preisser, John S., Andrzej T. Galecki, Kurt K. Lohman and Lynne E. Wagenknecht (2000), 'Analysis of smoking trends with incomplete longitudinal binary responses', *Journal of the American Statistical Association* **95**(452), 1021–1031. [16]

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org* [74]

Rao, J. N. K. and J. Shao (1992), 'Jackknife variance estimation with survey data under hot deck imputation', *Biometrika* **79**, 811–822. [25, 26]

Robins, James M., Andrea Rotnitzky and Lue Ping Zhao (1995), 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association* **90**, 106–121. [7, 16, 18, 20, 21, 22, 23, 24, 25, 31, 40, 104, 105]

Rotnitzky, Andrea, Christina A. Holcroft and James M. Robins (1997), 'Efficiency comparisons in multivariate multiple regression with missing outcomes', *Journal of Multivariate Analysis* **61**, 102–128. [16]

Rotnitzky, Andrea, James M. Robins and Daniel O. Scharfstein (1998), 'Semiparametric regression for repeated outcomes with nonignorable nonresponse', *Journal of the American Statistical Association* **93**, 1321–1339. [16, 20, 23]

Rubin-Bleuer, Susana and Ioana Schiopu Kratina (2005), 'On the two-phase framework for joint model and design-based inference', *The Annals of Statistics* **33**(6), 2789–2810. [29]

Rubin, Donald B. (1976), 'Inference and missing data', *Biometrika* **63**, 581–592. [15]

Rubin, Donald B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley series in probability and mathematical statistics, Wiley, New York. [25]

Särndal, Carl-Erik (1992), 'Methods for estimating the precision of survey estimates when imputation has been used', *Survey Methodology* **18**(2), 241–252. [46, 48, 86]

Särndal, Carl-Erik, Bengt Swensson and Jan Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York. [32, 48, 59]

Scott, Alastair and T. M. F. Smith (1974), 'Linear superpopulation models in survey practice', *The Indian Journal of Statistics, Series C* **36**(3), 143–146. [27]

Shao, Jun (2003), *Mathematical Statistics*, 2 edn, Springer-Verlag, New York. [29, 32]

Shao, Jun and Bob Zhong (2006), 'On the treatment effect in clinical trials with dropout', *Journal of Biopharmaceutical Statistics* **16**(1), 25–33. [25]

Shao, Jun and Philip Steel (1999), 'Variance estimation for survey data with composite imputation and nonnegligible sampling fractions', *Journal of the American Statistical Association* **94**, 254–265. [38, 43]

Song, Peter X.-K. (2007), *Correlated Data Analysis: Modeling, Analytics, and Applications*, Springer Series in Statistics, Springer, New York. [2, 9, 69]

Song, Peter Xue-Kun (1997), 'Generating dependent random numbers with given correlations and margins from exponential dispersion models', *Journal of Statistical Computation and Simulation* **56**, 317–335. [71]

Song, Peter Xue-Kun (2000), 'Multivariate dispersion models generated from gaussian copula', *Scandinavian Journal of Statistics* **27**(2), 305–320. [71]

Statistics Canada (1995), *Microdata User Guide, National Longitudinal Survey of Children and Youth - Cycle 1*. September 1994 to May 1995. [5]

Statistics Canada (2003), *Survey methods and practices*, Statistics Canada. Catalogue no. 12-587-XPE. [16]

Statistics Canada (2005), National longitudinal survey of children and youth: Home environment, income and child behaviour, The Daily, Statistics Canada. Catalogue number 11-001-XIE. [4]

Taylor, Jeremy M. G., Kristine L. Cooper, John T. Wei, Aruna V. Sarma and Trivellore E. Raghunathan (2002), 'Practice of epidemiology. Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African-American men', *American Journal of Epidemiology* **156**(8), 774–782. [25]

Thomas, Eleanor M. (2004), Aggressive behaviour outcomes for young children: Change in parenting environment predicts change in behaviour, Children and Youth Research Paper Series, Statistics Canada. Catalogue number 89-599-MIE. [3]

Víšek, Jan Ámos (1979), Asymptotic distribution of simple estimate for rejective, Sampford and successive sampling, *in* 'Contributions to Statistics', Reidel, Dordrecht, pp. 263–275. [32]

Wang, Qihua, Oliver Linton and Wolfgang Härdle (2004), 'Semiparametric regression analysis with missing response at random', *Journal of the American Statistical Association* **99**(466), 334–345. [16]

Wu, Changbao (2003), 'Optimal calibration estimators in survey sampling', *Biometrika* **90**(4), 937–951. [28]

Wu, Changbao and J. N. K. Rao (2006), 'Pseudo-empirical likelihood ratio confidence intervals for complex surveys', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **34**(3), 359–375. [32]

Yi, Grace Y. and Mary E. Thompson (2005), 'Marginal and association regression models for longitudinal binary data with drop-outs: A likelihood-based approach', *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **33**(1), 3–20. [16, 26]

Yuan, Ke-Hai and Robert I. Jennrich (1998), 'Asymptotics of estimating equations under natural conditions', *Journal of Multivariate Analysis* **65**, 245–260. [29]