TO WEIGHT OR TO ADJUST: AN EMPIRICAL STUDY OF THE DESIGN-BASED AND MODEL-BASED APPROACHES

Tianji Cai

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Sociology

Chapel Hill
2010

Approved by:
GUANG GUO
KATHLEEN MULAN HARRIS
FRANCOIS NIELSEN
WILLIAM D. KALSBEEK
SHARON L.CHRIST

ABSTRACT

Tianji Cai: To Weight or To Adjust: An empirical Study of the Design-based and Model-based Approaches

(Under the direction of Guang Guo)

When a sampling design is correlated to the dependent variable, then the distribution of the sampled units is different from that obtained from a simple random sampling design. Then the sampling design is informative, in the sense that if the design variables were not included in the analysis model, even conditional on the covariates, the estimated model parameters can be biased.

Questions have been asked about how survey data are modeled when sampling designs are informative. Two fundamental methodologies, design-based and model-based, have been proposed to address this issue. A model-based method--so-called sample distribution method, has been proposed by Krieger and Pfeffermann (1992; 1997) to extract the model of the sample data as a function of the model holding in the population and the sampling design. Once the model holding in the sample data is derived, the standard model-based analysis techniques can be applied to estimate the unknown population parameters. The core topic of this dissertation is to assess various modeling strategies and estimators of regression coefficients and their variance—both design-based and model-based, in particular, the sample distribution method, under the informative sampling design, and to develop a modeling strategy for analysts who are facing this design-based or model-based dilemma.

The dissertation is comprised of three research papers that provide 1) an evaluation of the design-based and model-based estimators under a single-stage informative sampling design; 2) an assessment of design-based and model-based estimators under an informative two-stage clustering sampling design; 3) a joint treatment of informative sampling and unit dropouts in longitudinal studies.

When a single-stage sampling design is informative, the model-based naïve method—either ordinary least square or maximum likelihood, produces biased results. The design-based method reduces the amount of biases for some parameters (e.g. intercept) but increases variances, which may lead to too conservative conclusions. The sample distribution method produces better estimates in the term of having smaller biases and variances than the naïve and design-based methods.

Under an informative two-stage clustering sampling design, ignoring the sampling effect, the model-based naïve method produces biased results. Under some specific assumptions, , the sample distribution method produces better estimators in terms of smaller biases and higher coverage rates compared to the naïve method and the design-based multilevel pseudo likelihood method. Although many previous studies have shown that multilevel pseudo likelihood method is preferred to compensate for the sampling design, this study shows that a rather simpler method—the sample distribution method can be used to address the design effect.

In a specific statistical setting, the relative performance of the design-based and the model-based methods for compensating the informative sampling design and dropout has been investigated. The simulation results indicate that both the model-based and the design-based approaches generally work well in the missing at random and missing not at random settings. Moreover, the sample distribution method combined with the Diggle and Kenward model has advantages of correcting the design effect and the nonignorable dropout.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Classical statistical models assume that any data being analyzed are from simple random samples. Unfortunately, in most studies, this is not the case. In fact, using sample random sampling in large-scale surveys is rare. Usually, sampling techniques such as unequal probability selection, stratification, and/or clustering are implemented to save money and time.

Questions have been asked (see e.g. Skinner et al 1989) about how survey data are modeled when sampling designs introduce dependencies and generate non-representative samples compared to the population. Two fundamental methodologies, design-based and model-based, have been proposed (see e.g. Lehtonen, R. and E. Pahkinen, 2004) for use with survey data. The key difference between these two methodologies is the source of variability. For the design-based method, the population from which the sample is selected is considered fixed; the variability of estimated parameters is from the difference in the samples. Therefore, if a census were taken, there would be no variability for the parameters are of interest. On the other hand, for the model-based method, the variability of parameters is from the stochastic mechanism that generates the data and the sampling design. If a census were taken, the variability of parameters of interest would still exist. Another key difference is the target of inference. The design-based method treats the target of inference as a fixed finite population. The inference is done over estimates from all possible samples about finite population quantities, and the sampling information is included. The model-based method requires the specification of a model that generates an outcome variable; the model is used to predict non-sampled values, and finite population quantities. The model is used in order to draw inferences on the superpopulation, which is more general than the finite population from which the sample is drawn. The model includes sampling information only when it is non-ignorable or part of the mechanism that generates data.

Little (2004) mentioned that there is a so-called “inferential schizophrenia” when researchers have to choose the most appropriate approach or when two approaches give different answers for scientific questions. Many studies have tried to address this issue (e.g., Asparouhov 2006). However, it is still unclear which approach is appropriate under given circumstances since the performance of an approach depends on actual survey design and data features (Asparouhov, 2006). The core topic of my PhD dissertation is to assess various modeling strategies and estimators of regression coefficients and their variance—both design-based and model-based under the informative sampling design, and to develop a modeling strategy for analysts who are facing this design-based or model-based dilemma.

The standard model-based analysis of survey data often ignores the complex nature of the sampling design, such as unequal probability selection, stratification, clustering, and nonrandom dropout. Although it is possible to incorporate all the design variables in the analysis model, it becomes impracticable since either the number of the design variables in a survey is large, or the design variables are not of interest. However, when the sampling design is correlated to the dependent variable, then the distribution of the sampled units is different from that obtained from a simple random sampling design. Then the sampling design is informative, in the sense that if the design variables were not included in the analysis model, even conditional on the covariates, the estimated model parameters can be biased. If the sampling design is not correlated to the response variable, excluding the design variables will not lead to biased estimates.

Krieger and Pfeffermann (1992; 1997) proposed a so-called sample distribution method to extract the model of the sample data as a function of the model holding in the population and the sampling design. Once the model holding in the sample data is derived, the standard model-based analysis techniques can be applied to estimate the unknown population parameters.

In Chapter 2, I introduce the design-based and model-based estimators, particularly the sample distribution estimator, and discuss the relative performance of design-based and model-based approaches under single-stage sampling designs.

The performance of design-based and model-based approaches under a multistage informative sampling design is further investigated in Chapter 3. Both the design-based and the model-based, in particular, the sample distribution estimators are derived under two typical multistage sampling designs. The first one is a cross sectional 2-stage design and the other one is a longitudinal design. According to the designs, two statistical models are estimated: cross-sectional multilevel linear model, and longitudinal model with first order autoregressive structure.

Chapter 4 extends the methods developed in Chapter 3 to missing data or the dropout problem. Comparison is conducted between the design-based model using weights and the model-based methods.

Chapter 5 is an empirical application. Using Add Health data as example, a model from a published paper is reexamined under various modeling strategies. Both design-based and model-based results are presented and discussed.

CHAPTER 2

GENERALIZED LINEAR MODEL UNDER A SINGLE-STAGE INFORMATIVE SAMPLING DESIGN

2.1. Introduction

In survey data analysis, weights are generated to reflect unequal sample probabilities of inclusion and to compensate for nonignorable nonresponse and frame undercoverage. However, the role of weights in regression is a subject of controversy. It has been well documented that if the parameter of interest is a finite population quantity such as mean, sum, ratio etc., then sampling weights must be used to correct deign effect to make inference. However, the role of sampling weights in regression has been debated extensively in the literature (e.g. Brewer and Mellor, 1973; Little, 1993; Pfeffermann, 1993).

Section 2 contains a general introduction of design-based and model-based approaches. Section 3 defines both the design-based and model-based estimators that are used in this study, and outlines several measures of the informativeness of a sampling design. Section 4 presents a simulation study. The last section summarizes the simulation results with discussions.

2.2. Background

Nearly all quantitative research in social science involves analysis of survey data which were collected using probability-based, rather than simple random design. Increasingly, questions have been asked about how survey data should be modeled when sampling designs introduce dependencies and generate non-representative samples compared to the population. Two approaches—model-based and design-based, have been proposed and implemented to analyze survey data.

2.2.1 Model-based approach

The model-based approach assumes that a model generates the data. If the theoretical model is true, then the sample design should have no effect as long as the probability of selection is not related to the dependent variable. Thereby, the use of weights is not necessary since they inflate the variance estimates if the proposed model between the response and the covariates is correct. Using weights also complicates the interpretation of the results (Hoem, 1989; Fienberg, 1989; Mislevy and Sheehan, 1989; Pfeffermann, 1996). Pfeffermann (1996) showed the simplest example to estimate the mean of a population. Suppose a sample *S* with *n* units was selected, each of them was chosen independently with probability π_i from a normally distributed population with mean μ and variance σ^2. The model-based estimator for population mean μ is

$$\overline{Y}_m = \sum Y_i / n$$

with variance $Var(\overline{Y}_m \mid S) = \sigma^2 / n$; while the design-based weighted estimator is

$$\overline{Y}_d = \sum \omega_i Y_i / \sum \omega_i$$

where $\omega_i = 1/\pi_i$ with variance $Var(\overline{Y}_d \mid S) = \sigma^2 [\sum \omega_i^2 /(\sum \omega_i)^2]$,

which is greater than $Var(\overline{Y}_m \mid S) = \sigma^2 / n$ unless ω_i is a constant.

In addition, since weights under very few circumstances were simple inversed probabilities of inclusion, sampling weights are generated to adjust for nonresponse, frame uncoverage, missing data, and poststratification (Korn and Graubard, 1999). Many of these adjustments impose assumptions which may or may not be appropriate for the model.

Sometimes weights can be used to check if there are any misspecifications of the model. Korn and Graubard (1995) illustrated that when the model was misspecified, weighted and unweighted results could be very different. If the model is correct, weighted analysis should provide similar results to the unweighted analysis (Lohr and Liu, 1994). Several tests have been developed to checkk if sampling weights could be omitted (DuMouchel and Duncan, 1983; Fuller, 1984; Nordberg, 1989).

When the probability of inclusion is related to the dependent variable, excluding the sampling information in the model may lead to biased estimates (Korn and Graubard, 1995). Suppose Y is an outcome variable, Z is a set of design variables, and X is a set of independent variables. In general, if Y depends on both X and Z, to estimate the effects of X on Y, one also needs to estimate the relationship between X and Z. Otherwise, the estimated effect of X on Y could be biased (Graubard and Korn, 2002). When the probability of inclusion is independent of the dependent variable, and the design variables that determine the probability of inclusion have been included in the model, the standard inference procedures apply. It should be noted that including all interactions may lead to unstable estimates (Cook and Gelman, 2006). One may only include some of the design variables or allow parameters to vary according to the combination of design variables (Gelman, 2007).

2.2.2 Design-based approach

For the design-based approach, weights are needed to account for sample design for estimating finite population quantities as well as the regression coefficients. Weights could also be used to reduce the effect of informative sampling. Weighted estimates are more robust than unweighted ones when some independent variables are left out of the model.

Some researchers argue that weights can add robustness to the model by reducing dependence on model assumptions (Kalton, 1989; Thomas and Cyr, 2002; Patterson et al., 2002; Vermunt and Magidson, 2007; DuMouchel and Duncan, 1983; Pfeffermann and Holmes, 1985). For example, if the proposed regression model does not hold, or the assumptions of the model are not approximately satisfied, the inference and predicted values are not correct (Kish and Frankel, 1974). However, the weighted estimates still have meaningful interpretation (so-called design consistency) and are the best approximations of the model parameters under a given distance function (Pfeffermann, 1993). The weighted estimators work better, on average, than unweighted ones in predicting unobserved population values when some independent variables have been omitted (Pfeffermann, 1996). It should be noted that the inference based on weighted estimation is limited to populations that are similarly structured as the population where the samples are drawn (Kalton, 1983). The inference also assumes asymptotic normality, which may not be true when the sample size is small.

However, the benefits of weighted estimators do not come without a price. As illustrated above, the design-based variance is greater than the model-based variance, which means that the inference drawn may be more conservative than it should be.

In addition to the question of whether weights should be used, another issue that researchers have widely discussed is how weights should be incorporated into analyses. Many estimators have been proposed in the literature (see Pfeffermann, 1993; 1996 for a good review). In this chapter, I only focus on two popular ones. The first one is design-based, based on Pseudo-likelihood approach. The other is model-based, derived from weighted distributions so called the sample distribution estimators (Pfeffermann et al., 2006).

2.3. Parameter estimation

To access the effect of sampling design on the estimation of the unknown regression coefficents, this study evaluates two types of estimators that are popular in the literature.

2.3.1 Design-based estimator

For the traditional design-based estimator, the sample data is weighted in inverse proportion to sample inclusion probabilities when the population regression function is fitted.

Define the population log likelihood function

$$l(\theta \mid Y) = \sum_{i=1}^{N} \log f(y_i,\ \theta),$$

for the population density function (pdf) $f(y,\theta)$, where θ is the vector of parameters. The pseudo-likelihood estimator can be solved by maximizing the weighted sample likelihood

$$l(\theta \mid Y_S) = \sum_{i=1}^{n} \omega_i \log f(y_i,\ \theta),$$

where $\omega_i = 1/\pi_i$ which is the individual probability of inclusion, and i=1 to n.

Variance can be obtained by three ways:

1. *The Taylor series linearization estimato*r (Woodruff 1971). This estimator is derived from linear approximation. It is computationally efficient and has become a gold standard of variance estimation. The estimator is as follows,

$$V(\hat{B}) = (X'WX)^{-1} V\left[\sum_{i\in s} w_i x_i'(y_i - x_i'\hat{B})\right](X'WX)^{-1}.$$

Suppose we fit a linear model with only one covariate,

$$Y = \beta_0 + \beta_1 x + e$$

If we ignore finite population correction, the Taylor series linearization estimator would be

$$\hat{V}(\hat{B}_1) = \frac{n\sum_{i\in s}(x_i-\bar{x})^2(y_i-\hat{B}_0-\hat{B}_1x_i)^2}{(n-1)\left[\sum_{i\in s}(x_i-\bar{x})^2\right]^2}$$

which is different from the model-based estimator as below even when simple random sample was drawn.

$$\hat{V}(\hat{\beta}_1) = \frac{\sum_{i\in s}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2}{(n-2)\sum_{i\in s}(x_i-\bar{x})^2} \quad \text{(Lohr, 1999)}$$

2. *Balanced repeated replication estimator* (BRR) (Kish and Frankel, 1974). It estimates variance via resampling techniques. The most common form of this estimator is, suppose the number of strata is large and with two PSUs in each stratum, a replicated sample is chosen by deleting one PSU per stratum and doubling the original weight of the remaining PSU. The deletion is assigned according to a corresponding Hadamard matrix. The variance of the estimator is obtained by

$$\hat{V}(\hat{B}) = R^{-1}\sum_{r=1}^{R}(\hat{B}_r-\hat{B})^2$$

where R is the total number of replicates; $\hat{B}$ is the full sample estimator; $\hat{B}_r$ is the estimated $\hat{B}$ for *r*-th replicate. A modified BRR estimator is available when one or more replicate estimators are undefined, while the full sample estimator is defined. See Rao and Shao (1999) for more information.

3. *Jackknife repeated replication estimator* (e.g. Wolter, 1985). This method is similar to the BRR in that a small and different portion of the total sample for each stratum or cluster is deleted to generate replicates. For example, delete-1 jackknife means that one PSU at time is deleted. Then the observation weights are adjusted

within the stratum where the PSU is deleted. This process is repeated for each stratum independently. The variance estimator is obtained by

$$\hat{V}(\hat{B}) = \sum_{r=1}^{R} \frac{m_r - 1}{m_r} (\hat{B}_r - \hat{B})^2$$

where m_r is the number of PSUs for *r*-th replicate; *R* is the total number replicates.

It has been well-documented that the standard error obtained from resampling variance estimators might be higher than the error from linearization estimators(e.g. Korn and Graubard (1999)). If the model is correctly specified in a simple random sample, one would arrive at a more conservative conclusion with the design-based approach. Although some researchers argue that the design-based approach protects against model misspecification, few of them illustrate the conditions where this conclusion may apply. The properties of design-based estimators are based on asymptotic conditions. The performance of design-based estimators is still unknown when the sample size is not large.

2.3.2 Sample distribution estimators

The second estimator discussed focused on in this study is based on the sample distribution approach. The sample distribution approach combines the knowledge of the method that is used to select the sample and population model by applying Bayes' theorem, to obtain the inference for the parameters of interest. For example, denote the population distribution of response variable *Y* as $f_p(y)$, and let *I* be the indicator of whether a population member is selected into the sample or not. Then the sample distribution of *Y* $f_s(y)$ can be written as

$$f_s(y) = \frac{\Pr(I = 1 | Y = y, X = x) f_p(y)}{\Pr(I = 1 | X = x)}$$

We can see that if the indicator variable *I* is independent of the response variable *Y*, then the sample distribution is equivalent to the population distribution.

Krieger and Pfeffermann (1992; 1997) proposed a method to extract the model of the sample data as a function of the model holding in the population and the sampling design. Denote the model hold in the sample as $f_s(y_i \mid x_i)$ which could be obtained via Bayes theorem as

$$f_s(y_i \mid x_i) = f(y_i \mid x_i, I_i = 1) = \frac{\Pr(I_i = 1 \mid y_i, x_i) \times f_p(y_i \mid x_i)}{\Pr(I_i = 1 \mid x_i)}$$

where I_i is the sample indicator for *i*th subject. In general, the probability of inclusion $\pi_i = \Pr(I_i = 1 \mid Y, Z)$ is not the same as $\Pr(I_i = 1 \mid y_i, x_i)$. However, Pfeffermann et al (1998) showed that

$$\Pr(I_i = 1 \mid y_i, x_i) = \int \Pr(I_i = 1 \mid y_i, x_i, \pi_i) f_p(\pi_i \mid y_i, x_i) d\pi_i = E_p(\pi_i \mid y_i, x_i)$$

Then it yields

$$f_s(y_i \mid x_i) = \frac{E_p(\pi_i \mid y_i, x_i) \times f_p(y_i \mid x_i)}{E_p(\pi_i \mid x_i)}.$$

If we can specify $E_p(\pi_i \mid y_i, x_i)$ and $E_p(\pi_i \mid x_i)$, the model holding in the sample could be specified, but the form of the expectations under the population is often unknown. Pfeffermann, and Sverchkov (1999) showed that those expectations can be identified and estimated from the sample data. Denote $\boldsymbol{w_i} = \frac{1}{\boldsymbol{\pi_i}}$ as the sampling weights for the *i*th individual. The following relationships hold,

$$E_p(y_i \mid x_i) = \frac{E_s(w_i y_i \mid x_i)}{E_s(w_i \mid x_i)}$$

$$E_p(\pi_i \mid y_i, x_i) = \frac{1}{E_s(w_i \mid y_i, x_i)}$$

$$E_p(\pi_i \mid x_i) = \frac{1}{E_s(w_i \mid x_i)}$$

Therefore,

$$f_s(y_i \mid x_i) = \frac{E_s(w_i \mid x_i) \times f_p(y_i \mid x_i)}{E_s(w_i \mid y_i, x_i)}$$

and

$$f_p(y_i \mid x_i) = \frac{E_s(w_i \mid y_i, x_i) \times f_s(y_i \mid x_i)}{E_s(w_i \mid x_i)}$$

Although the exact form of $E_p(\pi_i \mid y_i, x_i)$ usually is unknown, under some regularity conditions, it may be approximated by low order polynomials in terms of y_i and x_i, or by exponentials via the Taylor series approximation. For example, under polynomials approximation, suppose $\mathbf{x}_i = \left(x_{i1}, \quad ..., \quad x_{im}\right)$ is the vector of independent variables, we have,

$$E_p(\pi_i \mid y_i, x_i) \approx \sum_{j=0}^{J} A_j y_i^j + h\left(\mathbf{x}_i\right)$$

where $h\left(\mathbf{x}_i\right) = \sum_{p=1}^{m} \sum_{k=1}^{K_p} B_{kp} x_{ip}^k$, A_j, and B_{kp} are unknown parameters which can be estimated from the sample. Following this model, the sample distribution can be written as

$$f_s(y_i \mid \mathbf{x}_i) \approx \frac{\sum_{j=1}^{J} \left(A_j E^{(j)}\right) f_p^{(j)}(y_i \mid \mathbf{x}_i) + \left[A_0 + h\left(\mathbf{x}_i\right)\right] \times f_p(y_i \mid \mathbf{x}_i)}{\sum_{j=1}^{J} \left(A_j E^{(j)}\right) + \left[A_0 + h\left(\mathbf{x}_i\right)\right]}$$

where $E^{(j)} = E_p\left(Y_i^j \mid \mathbf{x}_i\right)$, and $f_p^{(j)}(y_i \mid \mathbf{x}_i) = \dfrac{y_i^j f_p(y_i \mid \mathbf{x}_i)}{E^{(j)}}$.

While under exponentials approximation, we have,

$$E_p(\pi_i \mid y_i, x_i) \approx \exp\left(\sum_{j=0}^{J} A_j y_i^j + h(\mathbf{x}_i)\right)$$

The corresponding sample distribution can be written as

$$f_s(y_i \mid \mathbf{x}_i) = \frac{\exp\left(\sum_{j=1}^{J}\left(A_j y_i^j\right)\right) \times f_p(y_i \mid \mathbf{x}_i)}{E_p\left[\exp\left(\sum_{j=1}^{J}\left(A_j y_i^j\right)\right) \middle| \mathbf{x}_i\right]}$$

Argued by Skinner (1994) and Pfeffermann etc (1998), the exponentials approximation is appealing in the common situation where the sampling is carried out in several stages.

Krieger and Pfeffermann (1997) proposed a two-stage method to estimate the population parameters and those parameters that index the sample distribution such as A_j, and B_{kp}. In the first step, the sample distribution coefficients can be estimated from the observed probabilities $\boldsymbol{\pi_i}$, applying either polynomials or exponentials approximations. For example,

Firstly, regress w_i against y_i, x_i to obtain an estimate of $E_s(w_i \mid y_i, x_i)$. Then obtain $E_p(\pi_i \mid \mathbf{x}_i)$ by integrating $E_s(w_i \mid y_i, \mathbf{x}_i)$ as follows,

$$E_p(\pi_i \mid \mathbf{x}_i) = \int E_p(\pi_i \mid y_i, \mathbf{x}_i) f_p\left(y_i \middle| \mathbf{x}_i\right) dy_i = \int \frac{1}{E_s(w_i \mid y_i, \mathbf{x}_i)} f_p\left(y_i \middle| \mathbf{x}_i\right) dy_i$$

and compute $E_s(w_i \mid \mathbf{x}_i) = \dfrac{1}{E_p(\pi_i \mid \mathbf{x}_i)}$. In the second step, the population parameters can be estimated by any standard method via substituting the estimated sample parameters.

Pfeffermann and Sverchkov (2003) considered two estimating equations for the sample distribution approach. The first one is directly derived from the sample distribution $f_s(y_i \mid \mathbf{x}_i)$. The first estimating equation is defined as

$$
\begin{aligned}
W_{1s}(\beta) &= \sum_s E_s\left(\frac{\partial \log\left(f_s(y_i \mid \mathbf{x}_i)\right)}{\partial \beta} \middle| \mathbf{x}_i\right) \\
&= \sum_s E_s\left(\frac{\partial}{\partial \beta}\log\left(\frac{E_p(\pi_i \mid y_i, \mathbf{x}_i)\times f_p(y_i \mid \mathbf{x}_i)}{E_p(\pi_i \mid \mathbf{x}_i)}\right)\middle| \mathbf{x}_i\right) \\
&= \sum_s E_s\left(\frac{\partial}{\partial \beta}\log E_p(\pi_i \mid y_i, \mathbf{x}_i) + \frac{\partial}{\partial \beta}\log_p(y_i \mid \mathbf{x}_i) - \frac{\partial}{\partial \beta}\log E_p(\pi_i \mid \mathbf{x}_i)\middle| \mathbf{x}_i\right) \\
&= \sum_s E_s\left(0 + \frac{\partial}{\partial \beta}\log_p(y_i \mid \mathbf{x}_i) + \frac{\partial}{\partial \beta}\log E_s(w_i \mid \mathbf{x}_i)\middle| \mathbf{x}_i\right) = 0
\end{aligned}
$$

$$
W_{1s}\left(\hat{\beta}\right) = \sum_s\left(\frac{\partial}{\partial \beta}\log\left(\frac{E_p(\pi_i \mid y_i, \mathbf{x}_i)\times f_p(y_i \mid \mathbf{x}_i)}{E_p(\pi_i \mid \mathbf{x}_i)}\right)\middle| \mathbf{x}_i\right) = 0
$$

The first estimator is defined as the solution of $W_{1s}\left(\beta\right)$.

For example, suppose the response variable y_i under study is normally distributed in population with mean $x_i'\beta$, and variance σ_e^2.

$$
y_i = \mathbf{x}_\mathbf{i}'\beta + e_i \quad e_i \sim N\left(0, \sigma_e^2\right)
$$

Under informative sampling, the sample inclusion probabilities have conditional expectation

$$
E_p(\pi_i \mid y_i, x_i) \approx \exp\left(A_0 + A_1 y_i + h\left(\mathbf{x}_i\right)\right).
$$

Then the sample distribution of response variable y_i is

$$
\begin{aligned}
f_s(y_i \mid \mathbf{x}_i) &= \frac{\exp\left(A_0 + A_1 y_i + h\left(\mathbf{x}_i\right)\right)\times \frac{1}{\sqrt{2\pi\sigma_e^2}}\exp\left(-\frac{\left(y_i - \mathbf{x}_\mathbf{i}'\beta\right)^2}{2\sigma_e^2}\right)}{\int \exp\left(A_0 + A_1 y_i + h\left(\mathbf{x}_i\right)\right)\times \frac{1}{\sqrt{2\pi\sigma_e^2}}\exp\left(-\frac{\left(y_i - \mathbf{x}_\mathbf{i}'\beta\right)^2}{2\sigma_e^2}\right)dy_i} \\
&= \frac{1}{\sqrt{2\pi\sigma_e^2}}\exp\left(-\frac{\left(y_i - A_1\sigma_e^2 - \mathbf{x}_\mathbf{i}'\beta\right)^2}{2\sigma_e^2}\right)
\end{aligned}
$$

Thus the linear model of y_i on $\mathbf{x}'$ in the sample is the same as the linear model in the population except the intercept is shifted by a constant $A_1\sigma_e^2$.

If the sample inclusion probabilities have conditional expectation

$E_p(\pi_i \mid y_i, x_i) \approx \exp\left(A_0 + A_1 y_i + A_2 y_i^2 + h\left(\mathbf{x}_i\right)\right)$,

the linear model of y_i on $\mathbf{x}'$ in the sample is different from the one that holds in the population by shifting the mean with a constant $A_1\sigma_e^2$, and new variance $\frac{\sigma_e^2}{1-2A_2\sigma_e^2}$.

$$f_s(y_i \mid \mathbf{x}_i) = \frac{\exp\left(A_0 + A_1 y_i + A_2 y_i^2 + h\left(\mathbf{x}_i\right)\right) \times \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_\mathbf{i}'\beta\right)^2}{2\sigma_e^2}\right)}{\int \exp\left(A_0 + A_1 y_i + A_2 y_i^2 + h\left(\mathbf{x}_i\right)\right) \times \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{\left(y_i - \mathbf{x}_\mathbf{i}'\beta\right)^2}{2\sigma_e^2}\right) dy_i}$$

$$= \frac{1}{\sqrt{2\pi\left(\frac{\sigma_e^2}{1-2A_2\sigma_e^2}\right)}} \exp\left(-\frac{\left(y_i - A_1\sigma_e^2 - \mathbf{x}_\mathbf{i}'\beta\right)^2}{2\left(\frac{\sigma_e^2}{1-2A_2\sigma_e^2}\right)}\right)$$

The solution of the sample likelihood function for the unknown parameters β yields the first sample distribution estimator $W_{1s}\left(\beta\right)$.

It has been showed that $E_p(y_i \mid x_i) = \frac{E_s(w_i y_i \mid x_i)}{E_s(w_i \mid x_i)}$, thus, the second estimating equation can be constructed as follows,

$$\begin{aligned} W_2\left(\beta\right) &= \sum_N E_p\left(\frac{\partial \log\left(f_p(y_i \mid \mathbf{x}_i)\right)}{\partial\beta} \middle| \mathbf{x}_i\right) \\ &= \sum_s \left(\frac{1}{E_s(w_i \mid \mathbf{x}_i)} E_s\left(w_i \frac{\partial \log\left(f_s(y_i \mid \mathbf{x}_i)\right)}{\partial\beta} \middle| \mathbf{x}_i\right)\right) \\ &= \sum_s E_s\left(\frac{w_i}{E_s(w_i \mid \mathbf{x}_i)} \frac{\partial \log\left(f_s(y_i \mid \mathbf{x}_i)\right)}{\partial\beta} \middle| \mathbf{x}_i\right) \\ &= \sum_s E_s\left(q_i \frac{\partial \log\left(f_s(y_i \mid \mathbf{x}_i)\right)}{\partial\beta} \middle| \mathbf{x}_i\right) = 0 \end{aligned}$$

where $q_i = \frac{w_i}{E_s(w_i \mid \mathbf{x}_i)}$.

$$W_{2s}\left(\hat{\beta}\right)=\sum_{s}\left(q_i\frac{\partial\log\left(f_s(y_i\,|\,\mathbf{x}_i)\right)}{\partial\beta}|\mathbf{x}_i\right)=0$$

Solving $W_{2s}\left(\hat{\beta}\right)$ yields the second estimator.

For the first estimator, a variance estimator can be obtained from the inverse of the information matrix evaluated at the estimator.

$$\widehat{\text{var}}\left(\hat{\beta}_{1e}\right)=\left[-E_s\left(\frac{\partial W_1(\beta)}{\partial\beta}\Big|_{\beta=\hat{\beta}_{1e}}\right)\right]^{-1}$$

While the variance of the second estimator can be estimated as

$$\widehat{\text{var}}\left(\hat{\beta}_{2e}\right)=\left(\frac{\partial W_2(\beta)}{\partial\beta}\Big|_{\beta=\hat{\beta}_{2e}}\right)^{-1}\left[\sum_{s}\left(q_i\left(\frac{\partial\log\left(f_s(y_i\,|\,\mathbf{x}_i)\right)}{\partial\beta}\Big|_{\beta=\hat{\beta}_{2e}}\right)\right)^{2}\right]\left(\frac{\partial W_2(\beta)}{\partial\beta}\Big|_{\beta=\hat{\beta}_{2e}}\right)^{-1}$$

Another plausible way to estimate variance is via bootstrapping, which in principle accounts for all the sources of variation, such as estimation of coefficients that index the sample distribution, and the form of $E_p(\pi_i\,|\,y_i,\mathbf{x}_i)$.

$$\widehat{\text{var}_{boot}}\left(\hat{\beta}_e\right)=\frac{\sum_{b=1}^{B}\left(\hat{\beta}_e^b-\overline{\beta}\right)^2}{B}$$

where $\overline{\beta}=\frac{\sum_{b=1}^{B}\hat{\beta}_e^b}{B}$

2.3.3 Test for the informativeness.

The informativeness of sampling design is not directly measurable, though several measures have been suggested by the literature. One idea is to compare the design-based estimate to the unweighted estimate. If the comparison leads to significantly different results, then the design-based one is recommended. For example, Dumouchel and Duncan (1983) proposed a test for the ignorability of the design. A test statistic λ which follows an *F* distribution with p and (n-2p) degrees of freedom equals to

$$\boldsymbol{\lambda} = \hat{\mathbf{D}}'\left[\hat{\mathbf{V}}(\hat{\mathbf{D}})\right]^{-1}\hat{\mathbf{D}}$$

where $\hat{\mathbf{D}}$ is the vector of differences between the estimators obtained by unweighted and weighted approaches, $\hat{\mathbf{D}} = \hat{\boldsymbol{\beta}}_{\mathbf{u}} - \hat{\boldsymbol{\beta}}_{\mathbf{w}}$, $\hat{\mathbf{V}}(\hat{\mathbf{D}})$ is an estimator of variance covariance matrix of $\hat{\mathbf{D}}$; p is the number of parameters, and n is the total number of observations. Pfeffermann (1993) constructed a test statistic that could be used to determine whether the sampling design is informative.

$$\boldsymbol{I} = (\hat{\theta}_w - \hat{\theta}_u)'\left[\hat{V}(\hat{\theta}_w) - \hat{V}(\hat{\theta}_u)\right]^{-1}(\hat{\theta}_w - \hat{\theta}_u) \sim \boldsymbol{\chi}(\mathrm{p})$$

where $\hat{\theta}_w$ and $\hat{\theta}_u$ are the parameter estimates of the weighted and unweighted analysis, and $\hat{V}(\hat{\theta}_w)$ and $\hat{V}(\hat{\theta}_u)$ are the estimated variances of those parameters. The test statistic follows a χ distribution with degrees of freedom p which equals to dimension of $\boldsymbol{\theta}$.

Another test constructed by Asparouhov (2004; 2006) follows a *t* distribution instead of a χ distribution.

$$\boldsymbol{I}(Y) = \frac{|\mu - \hat{\mu}_u|}{\sqrt{\upsilon_u}}$$

where μ is the population mean of response variable Y, $\hat{\mu}_u$ is the unweighted estimate of μ, and υ_u is the unweighted estimate of variance of Y. Since the population parameter μ is unknown, this measure could be approximated by

$$\boldsymbol{I}(Y) = \frac{|\hat{\mu}_W - \hat{\mu}_u|}{\sqrt{\upsilon_u}}$$

where $\hat{\mu}_W$ is the weighted estimate of μ.

However, the use of those statistics is limited. For example, it can only be applied to the point estimation. It does not apply to the actual question of whether the population model is equal to the sample model. Pfeffermann and Sverchkov (1999) proposed a test that directly addresses this question. Denote the residual term associated with unit i in population as,

$$\varepsilon_i = y_i - E_p(y_i \mid \mathbf{x}_i)$$

It is plausible to test the hypotheses of the form $E_p(\varepsilon_i^k) = E_s(\varepsilon_i^k)$, k=1,2.... Since

$$E_p(\varepsilon_i^k) = \frac{E_s(w_i \varepsilon_i^k)}{E_s(w_i)},$$

then an equivalent set of hypotheses is

$$H_{0k}: \ \mathrm{corr}\left(\varepsilon_i^k, w_i\right) = 0, \ k\text{=1,2...}$$

The test statistic is

$$FTS(k) = \frac{1}{2} \frac{\log\left(\dfrac{1+r_k}{1-r_k}\right)}{\widehat{SD_{boot}}}, \quad k\text{=1,2...}$$

where r_k is the empirical correlation $r_k = \widehat{\mathrm{corr}}\left(\hat{\varepsilon}_i^k, w_i\right)$, and $\widehat{SD_{boot}}$ is the bootstrap standard deviation of $\frac{1}{2}\log\left(\frac{1+r_k}{1-r_k}\right)$. The test statistic has an asymptotic normal distribution with mean zero. Another test which follows the similar idea is to simply regress w_i against $\hat{\varepsilon}_i^k$ and test if the coefficient equals to zero using a t test as follows.

$$\hat{\varepsilon}_i^k = \alpha w_i + \eta_i \quad \text{k=1,2, or 3}$$

$$\text{t=}\frac{\hat{\alpha}}{se(\hat{\alpha})} \sim \text{t}(1)$$

2.4. A Simulation study

In this section, the relative performance of design-based and sample distribution estimators is assessed by a simulation study. To test the effect of the informativess of a sampling design, a sample is selected from both the informative and non informative designs. Both linear and non-linear regression models are considered.

2.4.1 Simulation design

The true model is generated according to given population parameters such as regression coefficients, and population variance. Outcomes of non-informative sampling designs are a function of simulated variables, which are independent of design variables; outcomes of informative models are both simulated independent variables and design variables. For each linear or non-linear outcome, there are three predicators including continuous and categorical ones.

For each simulation, one design-based estimator using Pseudo Maximum Likelihood (PML) is estimated. For the sample distribution approach, two estimators are estimated according to estimation equation 1 and 2 mentioned above, respectively.

In this study, the informativeness of sampling designs is tested by the t test proposed by Pfeffermann and Sverchkov (1999). All simulation designs are generated using SAS 9.2. For the design-based estimator, PML is estimated using SAS survey procedures; the sample distribution estimators are estimated by SAS PROC NLP. The standard model-based (naïve ordinary least square) estimator is used as the reference. Table 2.1 summarizes all simulation designs.

For each of the following two super population models (2.1), and (2.2) with three covariates, five hundred finite populations are generated, each of size 1,000,000

$$Y_i = 10 + 1x_{1i} + 2x_{2i} + .5x_{3i} + e_i \tag{2.1}$$

$$P\left(Y_i = 1\right) = \text{logit}\left(1 + 1x_{1i} + 2x_{2i} + .5x_{3i} + e_i\right) \tag{2.2}$$

Where $e_i \sim N\left(0,\ 16\right)$ for $i = 1, \ldots,\ 1{,}000{,}000$. To see the two superpopualtion models in the context of a sociological study, the first model can be considered as a regression model for individual incomes, and the second model is a model to predict if a person's income is high or low. Explanatory variable x_1 follows a Bernoulli distribution with mean .5 (e.g. gender). x_2 is from a normal distribution $N(0,10)$ (e.g. centered years of schooling). x_3 follows a uniform distribution $U(15,35)$ (e.g. centered age).

A sample of size 3,000 is drawn from each finite population using a single-stage probability proportional to size systemic (PPS SYS) sampling design. The selection probability is related to a size variable *z*. To incorporate the informativeness of the sampling design, three designs are defined as follows.

$$z_i = \exp(6 + .5y_i + .25 * x_{3i} + u_i) \tag{2.3}$$

$$z_i = \exp(-2 + .25y_i - .005y_i^2 + .5x_{3i} + u_i) \tag{2.4}$$

$$z_i = \exp(6 + .25x_{3i} + u_i) \tag{2.5}$$

where $u_i \sim N(0,1)$. The sampling designs using (2.3) and (2.4) as its size variable is informative, while the one using (2.5) is non informative since the size variable is not related to the response variable. Totally 3,000 samples are selected according to two super population models. To compute the estimators based on the sample distribution approach, a two-step estimation is used. Firstly, the unknown parameters in $E_p(\pi_i \mid y_i, x_i)$ are estimated using the following relationship.

$$E_p(\pi_i \mid y_i, x_i) = \frac{1}{E_s(w_i \mid y_i, x_i)}$$

Then the estimated coefficients are substituted into the sample likelihood function to estimate the population parameters.

To make all estimators comparable, only bootstrapping variance estimates which are based on 500 bootstrapping replicates are reported for all estimated parameters.

2.4.2 Analysis

All simulation designs are implemented using SAS 9.2. Each of designs is replicated by 500 times. The quality of estimates is evaluated by using the empirical relative bias, the empirical mean square error, and coverage rates. The relative bias is defined as follows,

$$\mathrm{RBias}\left(\hat{\theta}\right)=\frac{1}{\theta}\left(\frac{1}{500}\sum_{1}^{500}\left(\hat{\theta}_i-\theta\right)\right)$$

The relative mean square error is calculated using the following formula,

$$\mathrm{RMSE}\left(\hat{\theta}\right)=\frac{1}{\theta}\left(\frac{1}{500}\sum_{1}^{500}\left(\hat{\theta}_i-\theta\right)^2\right)$$

The coverage rate is calculated as the proportion of true parameter that falls within the 95% confidence region for each of replicated samples. The Confidence region for estimated θ is constructed by using 5th and 95th bootstrapping quintiles.

2.5. Results and discussion

In the following the results of simulation are reported. The parameters of interest are regression coefficients and population variance (for the linear model).

2.5.1 Summary of simulation results

Table 2.2-2.4 contain the empirical relative biases and the empirical mean square error obtained under the exponential sampling (Equation. (2.3)), the exponential sampling with quadratic term (Equation. (2.4)), and non-informative sampling (Equation. (2.5)) for each parameter, respectively. Table 2.5-2.6 contain the equivalent information for logistic model. Table 2.7 summaries the results of test of informativeness.

The main findings from Tables 2.2-2.7 are as follows,

(1). Under exponential sampling with 1st order correlation with the response variable, the OLS estimator of intercept is biased. W1s method reduces this bias substantially. This reflects the effect of informative sampling, as mentioned in Section 2 — the mean of the sample distribution is different from the mean of the population distribution by a constant. The W2s estimator of intercept is also biased. Among all estimators, W1s estimator has the lowest relative biases and MSE in most cases. When the inclusion probabilities are related to the 1st order of response, the PML estimator has better coverage rates for all parameters except the population variance. Actually, the coverage rate for population variance estimated by PML is the worst among all estimators—the coverage rate is .490. In contrast, two sample distribution estimators have much better coverage rates for population variance; for example, the coverage rate is .750, and .846 for W1s, and W2s, respectively. In summary under exponential sampling with 1st order correlation with the response, the W1s estimator produces the least biased and most efficient estimates.

(2). Under exponential sampling with 2nd order correlation with the response variable, all estimates are biased for all estimators except W1s. The W1s estimator is almost unbiased for all parameters. This confirms the analytical result discussed in Section 3.2. The W1s estimator also has the lowest MSE. To summarize, the W1s estimator is best in terms of accuracy and efficiency.

(3). When the sampling design is non-informative, all estimators are unbiased. Among all estimators, PML estimator has the highest RMSE in most of cases. OLS estimator has the highest coverage rate and lowest RMSE in most of cases. This confirms the literature that when the sampling design is non informative, OLS estimator is the best.

(4). The t test for the informativeness performs well for the liner model. Under the two informative sampling designs, the test rejects non hypothesis 96% , and 84%

times at 95% confidence level, respectively. When the sampling design is non informative, the test only rejects 1.8% of non hypotheses at 95% confidence level. However, the test does not work well for the nonlinear case. Both deviance residual and Person residual are used to conduct the t test. The test using the deviance residual works well for the informative sampling design—it reports 66% of significant results. But it also rejects 28.2% of non hypotheses for the noninformative sampling design. The test is over sensitive to the informativeness. The test using Pearson residual is under sensitive to the informativeness — it only rejects 13.8% of non hypotheses when the sampling design is informative, but rejects 5.6% of non informative sampling design.

2.5.2 Test the sensitivity of the specification of conditional expectation

For the sample distribution estimator, a two-stage estimation is used to solve the unknown population parameters. Those parameters that index the sample distribution such as A_j, and B_{kp} are estimated at the first stage. Then the estimated sample distribution parameters are substituted into the likelihood function which is solved at the second stage. Since the sample is selected using the inclusion probabilities that have an exponential form such as Equation (2.3) to (2.5), therefore, an exponential approximation is assumed at the first stage. However, before those sample distribution parameters being estimated, researcher needs to determine whether the quadratic term of the response variable should be kept. If the quadratic term needs to be kept, the second order correlation is assumed. If not, the first order correlation is assumed. To investigate the sensitivity of choosing the first order or the second order correlation, a separate study is conducted. The true inclusion probabilities are proportional to the second order correlation with the response variable in exponential form. Firstly, unknown parameters in $E_p(\pi_i \mid y_i, x_i)$ are estimated using the quadratic form

$$E_s(\pi_i \mid y_i, x_i) = \exp\left(a_0 + a_1 y_i + a_2 y_i^2 + a_3 x_{3i}\right)$$

Secondly, a t test is conducted for the quadratic term by regressing w_i against $\hat{\varepsilon}_i^2$ to see if the coefficient equals to zero, where $\hat{\varepsilon}_i^2$ is the squared ordinary least square residual term by regressing independent variables against response variable. If so, the quadratic term will be dropped. The estimated coefficients are then substituted to the estimation of population parameters. If the coefficient is not zero, the quadratic term need to be kept and substituted in the later estimation. The results are reported in Table 2.8.

84% of tests reported significant quadratic term, which indicates that the test serves well to detect the correct form of correlation. Compared to the results reported in Table 4, it can be seen that for the estimator W1s, the relative biases and MSE are higher, and the coverage rates become slightly worse. This is because 16% of times the form of conditional expectation is not correctly specified. But the difference between Table 2.4 and 2.8 is not dramatic – it is still acceptable. This suggests that it is plausible to determine the form of conditional expectation based on the test result.

2.5.3 Conclusion

In this chapter three methods are discussed for estimating the parameters of the superpopulation when the sampling design is informative. The naïve method—either ordinary least square or maximum likelihood, produces biased results. The design-based method reduces the amount of biases for some parameters (e.g. intercept) but increases variances, which may lead to too conservative conclusions. The sample distribution method, in particular W1s, produces better estimates in the term of having smaller biases and variances than the naïve and design-based methods. But one needs some analytical and programming skill to implement, for example, the form of sample likelihood needs to be specified in PROC NLP. The estimator W2s is in a relatively simpler form, but it does not outperform than the design based estimator.

CHAPTER 3

MULTILEVEL MODELS UNDER A TWO-STAGE INFORMATIVE SAMPLING DESIGN

3.1. Introduction

One feature of the survey data is clustering. Under a two-stage sampling design, the population elements are grouped into clusters according to characteristics such as city blocks, schools, and hospitals. Before elements are selected, a subset of clusters called primary sampling units (PSUs) is selected first. Then elements are drawn from each of selected PSUs. It has been well documented that the cluster sampling provides less precision than other sampling methods such as the simple random sampling or the stratified sampling (Kish, 1965). However, the cost of the cluster sampling usually is lower than the other methods since it does not require a complete list of elements in the population which can be difficult, costly, or even impossible to construct. One characteristic of the cluster data is that elements within each of selected PSUs may be correlated. When the cluster data is modeled, such correlation has to be considered.

The purpose of this chapter is to estimate the regression coefficients and their variance of the superpopulation model under a two-stage sampling design. Section 2.1 and 2.2 introduce the design-based and the model-based approaches to estimate the parameters of interest. In order to compare these two approaches, Section 2.3 defines the corresponding point and variance estimators for each of approaches. Section 3 outlines a simulation study. The final section summarizes the simulation results and discusses the findings

3.2. Background

Both the design-based and the model-based approaches have been proposed in the literature to model the clustering effect. The model-based approach estimates either a population-averaged model that adjusts the standard error of the parameters, or a random effect model which incorporates the cluster-specified random effects. The design-based approach estimates the population likelihood function from the sample likelihood function, and then solves the estimated population likelihood function for the unknown parameters. The variance for each of the estimated parameters is obtained by Taylor linearization or resampling techniques which use formulas specialized for the sampling design.

3.2.1 Model-based approach

Robust Standard Error

For the model-based approach, various ways of incorporating the clustering effect have been proposed in the literature. When the individuals within a cluster are correlated, the robust standard error has been routinely used to reduce the bias that is caused by the correlation. Eicker (1967) and Huber (1967) first introduced the robust standard error which is also called as the sandwich estimator. White (1980), Liang and Zeger (1986), and many others (Arellano,1987; Royall, 1986; Lin and Wei, 1989, to name a few) explicated and extended this estimator to a more general context. Since the variances of the estimated parameters equal to

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \mathbf{(X'X)^{-1}X'}\,\mathrm{var}(\mathbf{Y}\,|\,\mathbf{X})\mathbf{X(X'X)^{-1}} \quad ,$$

where $\hat{\boldsymbol{\beta}}$ is the vector of the estimated regression coefficients. If the individuals within a cluster are correlated, the variance of Y conditional on Xs is not a diagonal matrix. Then the off-diagonal elements need to be estimated. Because there is no reason to assume that individuals in different clusters are correlated, the covariance in different clusters can be simply set to zero. The sandwich estimator is defined as

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \mathbf{(X'X)^{-1}X'\hat{\varepsilon}\hat{\varepsilon}'X(X'X)^{-1}}$$

with $\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}'$ as the estimated variance of Y conditional on Xs. Liang and Zeger (1986) derived a similar sandwich estimator under the generalized estimating equations.

Multilevel Model

Another way to incorporate the clustering effect is to add random effects. Laird and Ware (1982) outlined a general linear form of a mixed model (multilevel model).

$$\mathbf{Y_i = X_i\boldsymbol{\beta} + Z_i b_i + e_i}$$

where $\mathbf{Z_i}$ denotes n random-effect design matrix for the *i*th cluster, and $\mathbf{b_i}$ is the cluster-specified random effect. Introducing a cluster-specified random effect not only controls the correlation within clusters, but also corrects the denominator degrees of freedom for the number of clusters.

Searle et al. (1992) provided a detailed derivation of the maximum likelihood estimator. For example, a likelihood function for a linear mixed model $\mathbf{Y_i = X_i\boldsymbol{\beta} + Z_i b_i + e_i}$ is defined as

$$L(\mathbf{Y} \mid \mathbf{X},\mathbf{Z},\boldsymbol{\beta},\mathbf{D},\sigma_e^2) = \frac{\exp\left(-\tfrac{1}{2}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right)}{(2\pi)^{\frac{N}{2}}\,|\mathbf{V}|^{\frac{1}{2}}}$$

where $\boldsymbol{V} = \mathrm{var}(\boldsymbol{Y}) = \boldsymbol{ZDZ'} + \boldsymbol{\sigma_e^2 I}$, $\boldsymbol{D}$ denotes the covariance matrix for the random effect vector **bi**, and $\boldsymbol{\sigma_e^2}$ is the variance of the error vector **ei.** Then the log likelihood function can be written as

$$l = \log L(\mathbf{Y} \mid \mathbf{X},\mathbf{Z},\boldsymbol{\beta},\mathbf{D},\sigma_\varepsilon^2) = -\tfrac{1}{2}N\log(2\pi) - \tfrac{1}{2}\log|\mathbf{V}| - \tfrac{1}{2}\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$$

The maximum likelihood estimator for the unknown parameters can be obtained by taking the partial derivate of the function with respect to each of the unknown parameters, setting the resulting score function to zero, and solving for the unknown parameters.

It is not clear that under which circumstances researchers should choose the robust standard error or the mixed model to account for the clustering effect. One possible factor to be considered is the number of clusters. If the number of clusters is small, the robust standard error might be a good choice; while if the data is seriously unbalanced, the mixed model is superior. Actually, the clustered robust standard error is also available for the mixed model. The only difference between the robust standard error and the clustered robust standard error is the constant for the finite population correction. Another possible factor is the scientific question that the researcher is interested in. If one is not interested in the cluster-specified random effect, then the robust standard error might be preferred.

Sample Distribution Method

The sample distribution approach has been extended to a two-stage sampling design. Pfeffermann, Moura, and Silva (2006) proposed a multi-level sample distribution approach for a two-stage informative sampling design. Firstly, a two-level model that holds in the sample data was extracted as a function of the assumed population model and the two stages of sample selection probabilities, Then the unknown population parameters were estimated using Bayesian methods.

Consider a two-level model for a response variable Y that holds in the population as follows

$$y_{ij} = \mu_i + x_{ij}^{'}\beta + \varepsilon_{ij} \;\; \left(\varepsilon_{ij} \sim N\left(0, \sigma_e^2\right), \text{ j=1,...,N}_i\right)$$
$$\mu_i = z_i^{'}\gamma + \eta_i \;\; \left(\eta_i \sim N\left(0, \sigma_u^2\right), \text{ i=1,...,M}\right)$$

Let $\theta_i = \left(\mu_i, \beta', \sigma_e^2\right)$ and $\lambda = \left(\gamma', \sigma_u^2\right)$ be two vectors that contain the level-1 and the e level-2 unknown parameters of the population model, respectively. The model that holds in the sample can be written as

$$f_{s_i}\left(y_{ij} \left| x_{ij}, \theta_i, \psi_1\right.\right) = \frac{E_p\left(\pi_{j|i} \left| y_{ij}, x_{ij}, \theta_i, \psi_1\right.\right) f_p\left(y_{ij} \left| x_{ij}, \theta_i\right.\right)}{E_p\left(\pi_{j|i} \left| x_{ij}, \theta_i, \psi_1\right.\right)}$$

$$f_s\left(\mu_i \left| z_i, \lambda, \psi_2\right.\right) = \frac{E_p\left(\pi_i \left| \mu_i, z_i, \lambda, \psi_2\right.\right) f_p\left(\mu_i \left| z_i, \lambda\right.\right)}{E_p\left(\pi_i \left| z_i, \lambda, \psi_2\right.\right)}$$

where *i* indicates the *i*th cluster in the sample, and ψ_1, and ψ_2 are additional possible unknown parameters, e.g. thresholds of a latent variable related to those expectations.

The joint sample density function can be written as

$$f\left(\mu_i, \beta', \sigma_e^2, \psi_1, \lambda, \psi_2, \sigma_u^2 \left| x_{ij}, z_i\right.\right) = \prod_{i \in s}\left[\prod_{j \in s_i} f_{s_i}\left(y_{ij} \left| x_{ij}, \mu_i, \beta', \sigma_e^2, \psi_1\right.\right)\right] f_s\left(\mu_i \left| z_i, \lambda, \psi_2, \sigma_u^2\right.\right)$$
$$\times p(\beta') p\left(\sigma_e^2\right) p\left(\sigma_u^2\right) p(\lambda) p(\psi_1) p(\psi_2)$$

which can be maximized by Markov Chain Monte Carlo (MCMC) algorithm.

Eideh and Nathan (2009) applied the same idea but used a different method to estimate the two-level model that holded in the sample data. Following Krieger and Pfeffermann (1992; 1997), the conditional sample distribution of the random effect μ_i is

$$f_s(\mu_i \mid \mathbf{z}_i) = \frac{E_p(\pi_i \mid \mu_i, \mathbf{z}_i) \times f_p(\mu_i \mid \mathbf{z}_i)}{E_p(\pi_i \mid \mathbf{z}_i)}$$

where $\mathbf{z}_i$ is the vector of predictors for random effect μ_i, $f_p(\mu_i \mid \mathbf{z}_i)$ is the population distribution of the random effect μ_i conditional on $\mathbf{z}_i$. Similarly, the conditional sample distribution of y_{ij} given $\mu_i, \mathbf{x}_{ij}$ is

$$f_s(y_{ij} \mid \mu_i, \mathbf{x}_{ij}) = \frac{E_p(\pi_{j|i} \mid \mu_i, \mathbf{x}_{ij}, y_{ij}) \times f_p(y_{ij} \mid \mu_i, \mathbf{x}_{ij})}{E_p(\pi_{j|i} \mid \mu_i, \mathbf{x}_{ij})}.$$

And the marginal distribution of vector $\mathbf{y}_i$ is given by

$$f_s(\mathbf{y}_i) = \int \prod_{j=1}^{m_i} f_s(y_{ij}) d\mu_i$$

Thus the full sample likelihood function can be written as

$$f_s = \prod_{i=1}^{m} f_s(\mathbf{y}_i)$$

which can be maximized by any standard procedures.

3.2.2 Design-based approach

For the design-based approach, the robust standard error technique has also been derived to account for the clustering effect (for example Kish and Frankel (1974), Fuller (1975), and Binder (1983) to name a few). There is a slight numerical difference between the model-based and the design-based robust standard errors, which is caused by a constant multiplier.

Multi-stage weighting is another option for the design-based approach. Traditionally, the sampling weight is a single level variable; however, it becomes more and more common that survey data is collected by multistage sampling designs. For instance, Add Health study collects data in two stages: in the 1st stage a stratified random sample of 80 high schools and 52 middle schools was drawn with unequal probabilities of selection, and then a sample of individuals was drawn from each selected school (Harris, 2008). Single level weights may not carry adequate information to correct for higher level unequal probabilities of inclusion (Pfeffermann, et al., 1998, 24). To incorporate the multiple stage sampling design, multilevel weights that account for each sampling stages have been proposed.

Single level weights usually are defined as $\omega_i = 1/\pi_i$ where π_i is the individual probability of inclusion. When a multi-stage sampling strategy is used, the single level weights are the product of weights of each of stages. For example, the probability of the *j*th observation in the *i*th cluster being selected equals to the probability of the *i*th cluster being selected multiplied by the conditional probability of the *j*th observation being selected conditional on the *i*th cluster being selected. The weight for the observation *j* in cluster *i* is defined as

$$w_{ij} = \frac{1}{\pi_i \pi_{j|i}} \quad ,$$

the weight for the *i*th cluster is,

$$w_i = \frac{1}{\pi_i},$$

and the conditional weight for the *j*th observation in cluster *i* is

$$w_{j|i} = \frac{1}{\pi_{j|i}}$$

where π_i denotes the probability of the *i*th cluster being selected and $\pi_{j|i}$ denotes the conditional probability of the *j*th observation being selected given the *i*th cluster being selected. Similarly, one can define higher order weights. For instance, w_{ijk} is the weight that the *k*th observation in cluster *j* and in stratum *i* is selected.

However, the way that sampling weights are incorporated in a multilevel model is not straightforward. As mentioned earlier, the pseudo maximum likelihood estimator can be solved by maximizing the weighted sample likelihood function. For the mixed model, two common used estimating approaches are summarized as follows, and they are different in the way that how the multilevel weights are inserted to replicate the sampled elements.

Multilevel Pseudo Maximum Likelihood

Rabe-Hesketh and Skrondal (2006) and Aspraouhov (2006) proposed a method called Multilevel Pseudo Maximum Likelihood (MPML) which directly derives the population likelihood function by weighting the sample likelihood function

$$L_s(\theta_1,\theta_2)=\prod_i\left(\int\left(\prod_j f\left(y_{ij}\left|\mathbf{x}_{ij},\mu_i,\theta_1\right.\right)^{w_{j|i}}\right)\phi\left(\mu_i\left|\mathbf{z}_i,\theta_2\right.\right)d\mu_i\right)^{w_i}$$

where θ_1 and θ_2 are finite population parameters for the fixed effects, while μ_i is the cluster-specific random effect. It can be maximized by many techniques. As a general estimator, MPML can be extended to a generalized multilevel model. The weighted likelihood function can be maximized via many algorithms such as EM-algorithm, the Quasi-Newton algorithm, and the Fisher scoring algorithm. Stata (GLLAMM) and Mplus have implemented this estimator.

Probability Weighted Iterative Generalized Least Square

Instead of weighting the sample likelihood function, Pfeffermann et al. (1998) proposed an approach called Probability Weighted Iterative Generalized Least Square (PWIGLS) wherein the weights are incorporated into the process of solving the likelihood function for the unknown parameters.

For a linear mixed model, the variance covariance matrix of the response variable $\mathbf{Y}$ is a block diagonal matrix by cluster, and the solution of the population likelihood function can be written as a sum across clusters. Suppose there are Q random effects in total. The solution can be written as the sum of population statistics as follows

$$\begin{aligned}\hat{\boldsymbol{\beta}}&=\left(\mathbf{X}^{\mathbf{T}}\mathbf{V}^{\mathbf{-1}}\mathbf{X}\right)^{-1}\left(\mathbf{X}^{\mathbf{T}}\mathbf{V}^{\mathbf{-1}}\mathbf{Y}\right)\\&=\left(\sum_i\mathbf{X}_i^{\mathbf{T}}\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_i\mathbf{X}_i^{\mathbf{T}}\mathbf{V}_i^{-1}\mathbf{X}_i\right)\end{aligned}$$

and

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \left(\mathbf{M}^{\mathbf{T}}\left(\mathbf{V}^{-1}\otimes\mathbf{V}^{-1}\right)\mathbf{M}\right)^{-1}\left(\mathbf{M}^{\mathbf{T}}\left(\mathbf{V}^{-1}\otimes\mathbf{V}^{-1}\right)vec\left(\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right)^{\mathbf{T}}\right)\right)\\ &= \left(\sum_i \mathbf{M}_i^{\mathbf{T}}\left(\mathbf{V}_i^{-1}\otimes\mathbf{V}_i^{-1}\right)\mathbf{M}_i\right)^{-1}\left(\sum_i \mathbf{M}_i^{\mathbf{T}}\left(\mathbf{V}_i^{-1}\otimes\mathbf{V}_i^{-1}\right)vec\left(\sum_i\left(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\right)\left(\mathbf{Y}_i-\mathbf{X}_i\boldsymbol{\beta}\right)^{\mathbf{T}}\right)\right)\end{aligned}$$

where

$$\mathbf{M} = \left(vec\left(\mathbf{Z}_1\mathbf{Z}_1^{\mathbf{T}}\right)\ vec\left(\mathbf{Z}_1\mathbf{Z}_2^{\mathbf{T}}+\mathbf{Z}_2\mathbf{Z}_1^{\mathbf{T}}\right)\ ....vec\left(\mathbf{Z}_Q\mathbf{Z}_Q^{\mathbf{T}}\right)\ vec\left(\mathbf{I}\right)\right) \text{ and}$$

$$\mathbf{M}_i = \left(vec\left(\mathbf{Z}_{1i}\mathbf{Z}_{1i}^{\mathbf{T}}\right)\ vec\left(\mathbf{Z}_{1i}\mathbf{Z}_{2i}^{\mathbf{T}}+\mathbf{Z}_{2i}\mathbf{Z}_{1i}^{\mathbf{T}}\right)\ ....vec\left(\mathbf{Z}_{Qi}\mathbf{Z}_{Qi}^{\mathbf{T}}\right)\ vec\left(\mathbf{I}_i\right)\right)$$

Therefore the estimated population likelihood from the sample data should be a function of the weighted linear statistics. At each step of iteration, PWIGLS replaces the population quantities by the weighted sample statistics.

For the discrete response, the response has to be transformed by using a Taylor series-based linearization such as Marginal Quasi-likelihood, or Penalized Quasi-likelihood into a continuous one, and then PWIGLS can be utilized. This method has been implemented in commercial packages such as, LISREL, and MLwiN.

3.2.3 Estimators

In order to compare these two approaches, in this chapter, only two estimators for the unknown population parameters, such as fixed regression coefficients, their variance, and variance of the random effect, are studied. The first one is the sample distribution estimator which is model-based. The other one is the multilevel pseudo likelihood estimator which is design-based.

Sample Distribution Estimator

Eideh and Nathan (2009) applied the idea of extracting the model of the sample data as a function of the model holding in the population and the sampling design to multilevel models. Consider the same two-level population model as above. Similar as in Chapter 2, the cluster level and individual level probabilities of inclusion can be approximated by low order polynomials or by exponentials via the Taylor series approximation in terms of $\mathbf{z}_i$, μ_i, and y_{ij}.

$$E_p(\pi_i \mid \mu_i, \mathbf{z}_i) \approx g(\mathbf{z}_i)\exp\left(\sum_{r=0}^{R} b_r \mu_i^r\right) \tag{a}$$

$$E_p(\pi_i \mid \mu_i, \mathbf{z}_i) \approx g(\mathbf{z}_i) + \sum_{j=0}^{J} a_j \mu_i^j \tag{b}$$

$$E_p\left(\pi_{j|i} \left| y_{ij}, \mathbf{x}_{ij}, \mu_i\right.\right) \approx k\left(\mathbf{x}_{ij}, \mu_i\right)\exp\left(\sum_{h=0}^{H} d_h y_{ij}^h\right) \tag{c}$$

$$E_p\left(\pi_{j|i} \left| y_{ij}, \mathbf{x}_{ij}, \mu_i\right.\right) \approx k\left(\mathbf{x}_{ij}, \mu_i\right) + \sum_{m=0}^{M} d_m y_{ij}^m \tag{d}$$

where $g(.)$ and $k(.)$ are known functions. For example, under the first order exponential approximation, the cluster level probabilities of inclusion can be written as

$$E_p(\pi_i \mid \mu_i, \mathbf{z}_i) \approx g(\mathbf{z}_i)\exp(b_1 \mu_i).$$

Thus the sample distribution of random effect μ_i can be derived as

$$f_s\left(\mu_i \left| \mathbf{z}_i\right.\right) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left(-\frac{\left(\mu_i - \left(b_1\sigma_\mu^2 + \mathbf{z}_i'\gamma\right)\right)^2}{2\sigma_\mu^2}\right)$$

If the individual level probabilities of inclusion can also be approximated by the first order exponential, the conditional sample distribution of individual elements can be written as

$$E_p\left(\pi_{j|i} \left| y_{ij}, \mathbf{x}_{ij}, \mu_i\right.\right) \approx k\left(\mathbf{x}_{ij}, \mu_i\right)\exp\left(d_1 y_{ij}\right)$$

$$f_{s_i}\left(y_{ij} \left| \mu_i, \mathbf{x}_{ij}\right.\right) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{\left(y_{ij} - \left(\mu_i + \mathbf{x}_{\mathbf{ij}}'\beta + d_1\sigma_e^2\right)\right)^2}{2\sigma_e^2}\right)$$

And then the marginal sample distribution of cluster *i*, $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{im_i}\right)$ is given by

$$f_s\left(\mathbf{y}_i\right) = \left(2\pi\right)^{-\frac{m_i}{2}} \left(\sigma_e^2\right)^{-\frac{(m_i-1)}{2}} \left(m_i\sigma_\mu^2 + \sigma_e^2\right)^{-\frac{1}{2}}$$
$$\times \exp\left(-\frac{1}{2\sigma_e^2}\sum_{j=1}^{m_i}\left(y_{ij} - \left(\mathbf{z}_i^{'}\gamma + \mathbf{x}_{\mathbf{ij}}^{'}\beta + b_1\sigma_\mu^2 + d_1\sigma_e^2\right)\right)^2\right)$$
$$\times \exp\left(\frac{\sigma_\mu^2}{2\sigma_e^2\left(m_i\sigma_\mu^2 + \sigma_e^2\right)}\left(\sum_{j=1}^{m_i}\left(y_{ij} - \left(\mathbf{z}_i^{'}\gamma + \mathbf{x}_{\mathbf{ij}}^{'}\beta + b_1\sigma_\mu^2 + d_1\sigma_e^2\right)\right)\right)^2\right)$$

Thus the full sample likelihood function can be written as

$$f_s = \prod_{i=1}^{m} f_s\left(\mathbf{y}_i\right)$$

It can be maximized by standard procedures.

Since the informativeness parameters b_1 and d_1 are both unknown, they need to be estimated first in order to estimate the superpopulation parameters β, σ_μ^2 and σ_e^2. A two-stage estimation procedure is proposed by Eideh (2008). Similar as in Chapter 2, the first step is to estimate b_1 and d_1 by regressing $-\log\left(\boldsymbol{w}_i\right)$ and $-\log\left(\boldsymbol{w}_{j|i}\right)$ against μ_i and y_{ij}, respectively. One problem is that μ_i is not observed though it can be measured by, for example, $\bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}$ which is the cluster mean. However, substituting μ_i by $\bar{y}_i$ is not an ideal solution, since for μ_i, $\bar{y}_i$ is a measure with error.

$$\bar{y}_i = \mu_i + h_i,$$

where h_i is a random variable with variance $\frac{\sigma_h^2}{m_i}$.

Assuming σ_h^2 is known, Fuller (1987, p105) proposed a way to estimate b_1. For simplicity's sake, in this study, $\bar{y}_i$ is used as a measure of μ_i, thus an OLS estimator of b_1 can be easily obtained by using standard procedures, such as SAS PROC REG. The second step is to substitute those estimates to the full sample likelihood function, and solve it for the unknown parameters β, σ_μ^2 and σ_e^2.

Design-based estimator

The design-based estimator considered here is based on the work of Rabe-Hesketh and Skrondal (2006) and Aspraouhov (2006).

Suppose we have a population with a hierarchical structure in which level-1 (individual) elements are clustered in level-2 (cluster) units. The population likelihood function of a general multilevel model with the response variable Y_{ij} and two levels predictors $\mathbf{X}_{ij}$ and $\mathbf{Z}_i$ can be written as

$$L(\theta_1,\theta_2)=\prod_i\left(\int\left(\prod_j f\left(Y_{ij}\left|\mathbf{X}_{ij},b_i,\theta_1\right.\right)\right)\phi\left(b_i\left|\mathbf{Z}_i,\theta_2\right.\right)db_i\right)$$

where θ_1 and θ_2 are parameters for the fixed effect, while b_i is the cluster-specific random effect. Let the density function of Y_{ij} be $f\left(Y_{ij}\left|\mathbf{X}_{ij},b_i,\theta_1\right.\right)$ and the density function of b_i be

$\phi\left(b_i\left|\mathbf{Z}_i,\theta_2\right.\right)$.

Rabe-Hesketh and Skrondal (2006) and Aspraouhov (2006) proposed a method called Multilevel Pseudo Maximum Likelihood (MPML) which directly estimates the population likelihood function by weighting the sample likelihood function,

$$L_s(\theta_1,\theta_2)=\prod_i\left(\int\left(\prod_j f\left(y_{ij}\left|\mathbf{x}_{ij},b_i,\theta_1\right.\right)^{w_{j|i}}\right)\phi\left(b_i\left|\mathbf{z}_i,\theta_2\right.\right)db_i\right)^{w_i}$$

It can be maximized by many techniques.

Variance estimators

The inverse of Fisher information matrix is used to estimate the variance of the sample distribution estimator. Holding b_1 and d_1 as fixed, the variance of estimates of superpopulation parameters $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}},\ \hat{\sigma}_{\mu}^{2},\ \hat{\sigma}_{e}^{2}\right)$ is given by,

$$\widehat{\operatorname{var}}\left(\hat{\boldsymbol{\theta}}\right) = \left[I\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1} = \left[-\frac{1}{n}\left(\frac{\partial^2 L(\boldsymbol{\theta})}{\partial'\boldsymbol{\theta}\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)\right]^{-1}$$

where $L\left(\boldsymbol{\theta}\right)$ is the likelihood function of $\boldsymbol{\theta}$.

An alternative to Fisher information matrix is to estimate the variance via bootstrapping as shown in Chapter 2. Mentioned by Pfeffermann, and Sverchkov (2004), the bootstrapping variance accounts for two sources of variations: one is due to estimate b_1, and d_1, and the other one is caused by estimating the unknown superpopulation parameters. The bootstrapping variance is not used here for computational consideration.

The variance of the design-based estimator is estimated by Taylor linearization,

$$\widehat{\operatorname{var}}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\mathbf{J}}^{-1}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\left(\mathbf{x}_{ij} - \hat{\tau}_i\hat{\bar{\mathbf{x}}}_i\right)\left(y_{ij} - \mathbf{x}_{ij}\hat{\boldsymbol{\beta}}\right)\right\}\hat{\mathbf{J}}^{-1}$$

where $\hat{\mathbf{J}} = \sum_{i=1}^{m}\mathbf{x}_i'\mathbf{W}_i\hat{\mathbf{D}}_i$, $\mathbf{W}_i = diag\left(w_{i1},\ldots,w_{in_i}\right)$, $\hat{\mathbf{D}}_i = \mathbf{x}_i' - \hat{\tau}_i\hat{\bar{\mathbf{x}}}_i$. $\hat{\bar{\mathbf{x}}}_i = \sum_{j=1}^{n_i} w_{j|i}\mathbf{x}_{ij} \Big/ \sum_{j=1}^{n_i} w_{j|i}$

$\hat{\tau}_i = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_e^2 / \hat{N}_i + \hat{\sigma}_u^2}$, and $\hat{N}_i = \sum_{j=1}^{n_i} w_{j|i}$.

3.3. Simulation Study

To assess the performance of estimators under various sampling designs, a simulation study is carried out. There are totally four designs divided by the informativeness of each design at the sampling stages. Table 3.1 summarizes all simulation designs.

Firstly, the true model is generated according to following population model.

$$Y_{ij} = 4 + 1x_{1ij} + 2x_{2ij} + .5x_{3ij} + \mu_i + e_{ij}$$

where $\mu_i \sim N(0,3)$ and $e_{ij} \sim N(0,9)$.

Secondly, a sample with 300 clusters, each of which has 10 individuals, is selected by a two-stage sampling design using PPS systemic method. The first stage probability of inclusion is proportional to a size variable z which is defined as below

$$z_i = \exp(4 + .5 * \mu_i) \tag{3.1}$$

$$z_i = \exp(4 + .5 * z_{1i}) \tag{3.2}$$

where $\mu_i \sim N(0,3)$ is the random effect for the *i*th cluster; $z_1 \sim N(0,1)$ is an exogenous variable. The second stage probability of inclusion within each of the selected clusters is proportional to the size variable z_{ij} defined as below.

$$z_{ij} = \exp(6 - .2y_{ij}) \tag{3.3}$$

$$z_{ij} = \exp(6 - .2x_{3ij}) \tag{3.4}$$

where y_{ij} and x_{3ij} are values of the response variable y and the predictor x_3 of the *j*th individual in the *i*th cluster. The sampling design using Equation (3.1) or (3.3) as its size variable is informative, while the one using Equation (3.2) or (3.4) is noninformative since the size variable is not related to the response variable. The combination of using those four size variables to select samples gives us four simulation designs: one is informative at both stages; two are informative only at one stage; and the last one is noninformative at both stages.

All simulation designs are generated using SAS 9.2. Each of simulation designs is repeated 500 times. All estimators are programmed in SAS using PROC IML, except MPML which is conduced by GLLAMM in STATA. The nonlinear optimization of a likelihood function for the sample distribution estimator is carried out by the Newton-Raphson ridge method, and the corresponding Hessian matrix is calculated by finite differences method using CALL NLPFDD in PROC IML.

3.4. Results and Discussion

Table 3.2 to 3.5 report the empirical relative bias, the empirical mean square error, and the coverage rates for each of the superpopulation parameters from the simulation.

The main findings for each of the tables are as follows,

(1) When both sampling stages are informative, the naïve method and the multilevel pseudo likelihood method produce biased estimates, while the sample distribution estimates are almost unbiased. The estimated intercept and variances from the naïve method and the multilevel pseudo likelihood method are biased. The sample distribution method reduces the biases substantially, although the estimated intercept is still biased, which is caused by using $\bar{y}_i$ as a measure of μ_i. The superior performance of the sample distribution method confirms the analytical results presented in Section 2. Except for the intercept and the variances, other estimated coefficients are unbiased with close-to-zero relative bias, small mean square errors, and high coverage rates for the naïve method and the sample distribution method. Fixed effects are close to unbiased for the multilevel pseudo likelihood method, but with higher mean square errors and lower coverage rates.

(2) When the sampling design is informative at the first stage (cluster level), some of the estimated parameters obtained from the naïve method and the multilevel pseudo likelihood method are biased. The naïve estimates for the intercept are

biased. The sample distribution method improves the estimation of the intercept dramatically, but the amount of biases is still not close to zero. Except the intercept, all the other coefficients are almost unbiased for both the naïve method and the sample distribution method. The multilevel pseudo likelihood method produces unbiased estimates for some fixed effects, but fails for the cluster-specified random effect and the intercept.

(3) When the sampling design is informative at the second stage, the naïve method yields the best estimates, although the estimated intercept is biased. The sample distribution method increases the biases instead of reducing it. This may be caused by the small negative informativeness parameter d_1 -- the average estimated d_1 is -0.03. Same as the previous design, the multilevel pseudo likelihood method does not performance well for the intercept and variance, but produces unbiased estimates for some fixed effects under this sampling design.

(4) When neither of sampling stages is informative, the naïve and the sample distribution methods yield the same estimates. The multilevel pseudo likelihood method still does not work well for the variance of the random effect, but other parameters are unbiased.

3.5. Conclusion

In this chapter, both design-based and model-based multilevel models under a two-stage sampling design are evaluated under some specific assumptions by a simulation study. Ignoring the informative sampling design, the naïve method produces biased results. Under a two-stage exponential sampling design, the sample distribution method produces better estimators in terms of smaller biases and higher coverage rates compared to the naïve method and the multilevel pseudo likelihood method. Many previous studies have shown that multilevel pseudo likelihood method is preferred to compensate for the sampling design. However, this study shows that a rather simpler method—the sample distribution method can be used to address the design effect. It needs to be noticed that the numerical behavior of the multilevel pseudo likelihood under the informative sampling design can be problematic. About 2% of estimations did not converge after 16,000 iterations for the informative designs.

There are some limitations with respect to the analysis and simulation design that may affect the accuracy of the results. For example, a rather simple form of the sampling design is assumed. For the simplicity's sake, the probabilities of inclusion are also assumed to have a specific exponential form. The sample-distribution method is at a bit of advantages, since the approximation of the conditional expectation of the probabilities of inclusion for the sample distribution estimator is specified correctly according to the true model. The effect of misspecification of the approximation is not discussed in this study. All of these limit the generability of this study.

CHAPTER 4

JOINT TREATMENT OF INFORMATIVE SAMPLING AND UNIT DROPOUTS IN LONGITUDINAL STUDIES

4.1 Introduction

Missing data is a common problem in social science research. For example, in a longitudinal survey, respondents may refuse to participate in the first wave of data collection or drop out in subsequent waves due to moving. Additionally, respondents may not answer all of the questions asked in the study questionnaire. The first type of missing data is called unit droupout, while the second type of missing data is called item nonresponse. It is well known that restricting analysis to complete cases (those with no missing values) may produce biased and, less efficient estimates (e.g. Little and Rubin (1987)). In addition to complete case analysis, many methods have been proposed to compensate for missing information, based on some assumption of the missing mechanism.

In this paper, only monotone unit dropout is discussed. In other words, if respondents drop out, then they can not be included in subsequent waves. Section 2 contains a introduction of basic assumptions of missing patterns. Section 3.1 and 3.2 introduce three popular methods to compensate dropout in longitudinal studies. Section 3.3 considers a joint treatment of informative sampling and dropout. Section 4 outlines a simulation study. The final section summarizes the simulation results and discusses the findings.

4.2 Missing data mechanism

In the literature of missing data analysis, researchers usually follow some assumptions of the distribution of missingness. These assumptions may not be appropriate since the mechanism behind missingness is somewhat unknown. For mathematical convenience, missingness is treated as a probabilistic phenomenon (Rubin, 1976). Consider a longitudinal survey where the interest is to study the relationship between a set of independent variables Xs and a response variable Y. Assume that some of the respondents drop out after the first wave. Suppose R is an indicator variable which takes the value of 1 if the respondent drops out, or 0 otherwise. The missing mechanism can be then expressed as a conditional probability of the distribution of R given Xs and Y.

$$\boldsymbol{Pr}\left(\boldsymbol{R} \mid \boldsymbol{X},\boldsymbol{Y}\right)$$

where R=1 or 0.

There are three basic missing patterns. If the unit dropout does not depend on Xs and Y, then this situation is called Missing Complete at Random (MCAR), which implies

$$\boldsymbol{Pr}\left(\boldsymbol{R}=1 \mid \boldsymbol{X},\boldsymbol{Y}\right)=\boldsymbol{c},$$

where c is a constant. Complete data analysis follows this assumption, where the complete data set is a random subset of original data. Excluding the respondents with missing values does not harm the representativeness of the original data set. In practice, not many examples satisfy this assumption (Little, 1992).

If the unit dropout only depends on Xs or Y from previous waves of but not on current Y, the situation is called Missing at Random (MAR). Compared to MCAR, it is a relatively weaker assumption. We assume the distribution of R only depends on X, as

$$\boldsymbol{Pr}(\boldsymbol{R}=1 \mid \boldsymbol{X}, \boldsymbol{Y})=\boldsymbol{f}(\boldsymbol{X})$$

Most statistical procedures require MAR. However, in general, there are no easy ways to test whether MAR holds in a data set without follow-up studies for unit dropout. In empirical researches, the appropriate form of $\boldsymbol{f}(\boldsymbol{X})$ may not be found due to the lack of information. For instance, some variables contributing to R are not observed in the study. It has been shown that MAR is still a reasonable assumption (Rubin, Stern and Vehovar, 1995).

The third missing data mechanism is called Missing Not at Random (MNAR) if the distribution of R depends on both Xs and the current value of Y.

$$\boldsymbol{Pr}(\boldsymbol{R}=1 \mid \boldsymbol{X}, \boldsymbol{Y})=\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{Y})$$

For example, suppose that there is a longitudinal study of the relationship between Social economic status and delinquent behavior. If respondents who have low SES or high delinquency are more likely to drop out than those who have high SES or low delinquency, then the missing data is MNAR. When MNAR is present, one could specify $\boldsymbol{Pr}(\boldsymbol{R}=1 \mid \boldsymbol{X}, \boldsymbol{Y})$, along with $\boldsymbol{Pr}(\boldsymbol{X}, \boldsymbol{Y})$ (Heckman 1976). Another option is to specify the model as a mixture of two parts: one for $\boldsymbol{Pr}(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{R}=1)$ and another for $\boldsymbol{Pr}(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{R}=0)$ (Rubin, 1974; 1977; Little, 1993). A similar limitation is that one may not have enough information to build $\boldsymbol{Pr}(\boldsymbol{R}=1 \mid \boldsymbol{X}, \boldsymbol{Y})$, or even have no information on $\boldsymbol{Pr}(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{R}=0)$ if neither X or Y is observed. No model proposed for R cannot be easily verified.

4.3 Missing data in longitudinal studies

Both design-based and model-based methods have been proposed to handle missing data. For design-based approaches, weights are generated and assigned to complete cases to resemble a finite population based on sampling information and available data. For model-based approaches, missing values can be imputed by one (single imputation) or more than one (multiple imputation) set of plausible values. Then a standard procedure is applied to the imputed data which is treated as complete data. Missing patterns can also be modeled for example, the Diggle and Kenward model (1994), using some distribution assumptions about dropouts.

4.3.1 Design-based approach

Cross-sectional and Panel weighting

There are several ways to construct weights to compensate for unit dropout in a longitudinal study (Holt, Elliot, 1991), weighting classes and the post-stratification are two common methods. The weighting class method partitions the sample into "weighting classes" (cells), according to some predefined variables. Then the final weight for each of individuals within each cell is calculated as the base weight (e.g. the reciprocal of the probability of inclusion) multiplied by the reciprocal of the response rate in that class. The post-stratification calibrates the weights to an external set of population counts (e.g. from a recent census).

These methods do not take into account longitudinal designs, where a repeatedly-observed individual has the same weight across waves. Biemer and Christ (2007; 2008) proposed alternative methods to allow the adjustment for the wave-specific dropout. The wave-specific weights equal to the wave-specific dropout adjustment multiplied by the base weights.

The nonresponse adjustment to the weight for entry into the study at wave *t* is given by

$$\lambda_{it} = \frac{1}{\pi_{it} \mid \pi_i}$$

where $\pi_{it} \mid \pi_i$ is the conditional probability for the *i*th individual response at wave t given the inclusion probability π_i for the *i*th individual. Then the wave-specific, cross-sectional weight for the *i*th observation at wave *t* is,

$$\boldsymbol{w}_{it} = \lambda_{it}\boldsymbol{w}_i = \frac{1}{\pi_{it} \mid \pi_i} \times \frac{1}{\pi_i}$$

where $\boldsymbol{w}_i$ is the base weight.

The wave-specific, cross-sectional weights can be converted to panel weights for repeated measures by assuming the dropout at wave *t* is independent of dropout at wave *t*+1 conditional on the model generating the nonresponse probabilities. The panel weight for the *i*th individual at wave *t* is given by

$$\boldsymbol{w}_{it} = \frac{1}{\pi_{i1} \mid \pi_i} \times \frac{1}{\pi_{i2} \mid \pi_i} \times \ldots \times \frac{1}{\pi_{i(t-1)} \mid \pi_i} \times \frac{1}{\pi_{it} \mid \pi_i} \times \frac{1}{\pi_i}$$

It is not necessary for the individual to be observed consecutively, and any combinations of wave data can be used to construct panel weights (Christ, 2008).

The cross-sectional and panel weights can be used under MNAR as long as the probability distribution of the dropout process being modeled.

4.3.2 Model-based approaches

Single and Multiple imputation

Weighting is effective in terms of removing the dropout bias, but it ignores partial information from subjects with incomplete data (Raghunathan, 2004). Besides weighting, an alternative approach is to impute the missing values in the data set. This approach has the advantage that after imputation standard statistical procedures can be used on the imputed data, and global-adjustment formulas exist to adjust estimated points and standard deviations for multiply imputed data (Rubin and Schenker, 1991;

Rubin, 1996). The Knowledge of the missing data mechanism can be easily incorporated to produce imputed values. Many imputation methods have been developed, such as the hot deck and the mean substitution. Most of these methods can be categorized into a multiple regression framework (Kalton, and Kasprzyk, 1986).

Let X be the variable of which some of observations are missing across waves, and Z be the set of observed auxiliary variables that are used as predictors of missing values of X. The imputed value of X can be represented by the following regression equation,

$$\hat{x}_{mi} = z\beta + \hat{e}_{mi}$$

where $\hat{x}_{mi}$ is the imputed value for *i*th individual with missing value on X; z is the vector of auxiliary variables; β is the regression coefficient vector; $\hat{e}_{mi}$ is the error term for *i*th individual which is, for example, normally distributed with mean 0, variance σ^2. This is called single imputation.

The single imputation fails to reflect the uncertainty of the imputed values, since the imputed values are from a guess rather than observed true values. Therefore, the single imputation underestimates the variation of the target parameters. A straightforward justification for the model above is to impute multiple times with a different $\hat{e}_{mi}$ at each time, where all $\hat{e}_{mi}$ are from the same distribution such as $N\left(0,\hat{\sigma}^2\right)$. This is called multiple imputation which has been implemented in many commercial packages. After imputation, standard statistical procedures can be applied to the imputed data sets. Since more than one value has been imputed in multiple imputation, the point estimates and their standard errors need to be adjusted to obtain a single estimate. Rubin (1987) derived a global formula for combining the results from multiple analyses. Suppose M data sets have been imputed, which means each missing value has been imputed M times, let e_l be the *l*th estimate and s_l be its standard error, then the multiple imputation estimate is the average,

$$\bar{e}_{MI} = \frac{1}{M}\sum_{l=1}^{M} e_l \quad ,$$

where *l*=1,…, M. and its the standard error is

$$s_{MI} = \sqrt{\bar{\mu}_M + \frac{M+1}{M} b_M} \quad ,$$

where

$$\bar{\mu}_M = \frac{1}{M}\sum_{l=1}^{M} s_l^2 \text{ and}$$

$$b_M = \frac{1}{M-1}\sum_{l=1}^{M}\left(e_l - \bar{e}_{MI}\right)^2 \quad .$$

This method is reasonable if the sample size is large (Rubin, 1987). It has been commonly used in item dropout but is not recommended for data with more than 50 percent missing information (Rubin, 2003). It should be noticed that the purpose of imputation is to obtain valid inferences. Therefore, though imputation is a way to recover missing data, the imputed values should not be viewed as individual values.

Though multiple imputation might be reasonable, it does not take into account the uncertainty of the parameters of the predictors for missing values (Rubin, 1987). A more thoughtful method mentioned by Rubin (1987), and Little, and Rubin (2002) is the Bayesian approach, where the missing values are drawn from the posterior predictive distribution of the missing observations, conditional on the observed data.

Diggle and Kenward model

Imputation may give unbiased marginal distributions, but may also distort the association between variables (Brick, and Kalton, 1996). Since maximum likelihood has been widely accepted as a general method to estimate the unknown parameters, one can actually model the missing pattern based on the available sample information without imputing any missing values.

Assume the vector of outcomes $\mathbf{Y}_i = \left(Y_{i1} \quad Y_{i2} \quad ... \quad Y_{it}\right)$ is of interest. Let D_i denote the variable which indicates the occasion at which dropout occurs. If $D_i > 1$, then $\mathbf{Y}_i$ is split into the $(D_i - 1)$ dimensional observed part $\mathbf{Y}_i^o$ and the $(t - D_i + 1)$ dimensional missing part $\mathbf{Y}_i^m$, where *t* is the total number of follow ups for a balanced longitudinal design. The contribution of the *i*th individual to the likelihood function, based on the observed part $\left(\mathbf{y}_i^o, d_i\right)$, is proportional to the marginal density function,

$$f\left(\mathbf{y}_i^o, d_i \middle| \theta, \psi\right) = \int f\left(\mathbf{y}_i, d_i \middle| \theta, \psi\right) d\mathbf{y}_i^m = \int f\left(\mathbf{y}_i \middle| \theta, \psi\right) f\left(d_i \middle| \mathbf{y}_i, \psi\right) d\mathbf{y}_i^m$$

where $f\left(\mathbf{y}_i, d_i \middle| \theta, \psi\right)$ is the model for $\mathbf{Y}_i$ with the unknown parameters θ, and $f\left(d_i \middle| \mathbf{y}_i, \psi\right)$ is the model for the dropout process with unknown parameters ψ.

Suppose $D_i = j$, for a monotone unit dropout case which means, all subsequent waves after the *j*th wave are all missing, the dropout model is given by

$$P\left(D_i = j \middle| \mathbf{y}_i, \psi\right) = P\left(D_i = j \middle| h_{ij}, y_{ij}, \psi\right)$$
$$= \begin{cases} P\left(D_i = j \middle| D_i \geq j, h_{ij}, y_{ij}, \psi\right) & j\text{=}2 \\ P\left(D_i = j \middle| D_i \geq j, h_{ij}, y_{ij}, \psi\right) \times \prod_{k=2}^{j-1}\left[1 - P\left(D_i = k \middle| D_i \geq k, h_{ik}, y_{ik}, \psi\right)\right] & j = 3,...,t \\ \prod_{k=2}^{t}\left[1 - P\left(D_i = k \middle| D_i \geq k, h_{ik}, y_{ik}, \psi\right)\right] & j = t+1 \end{cases}$$

where $h_{ij} = \left(y_{i1}, y_{i2}, ..., y_{i;j-1}\right)$ is the observed history of the *i*th individual right before dropout time *j*. when *j*=2, an individual is only observed at the first wave. When *j*=*t*+1 there is no dropout. If the dropout happens after the second wave, and before the last follow-up, $j = 3,...,t$, the dropout model is a product of two components: the first part is corresponding to the probability of the dropout happening at wave j, and the second part is the joint probability of no dropout happened occurring before wave j.

For continuous outcomes, Diggle and Kenward (1994) proposed a logistic model for the dropout process. For example, assume the dropout process only depends on the previous and current values of the response variable, a Diggle and Kenward model can be written as

$$\operatorname{logit}\left(P\left(D_i = j \middle| \mathbf{y}_i, \psi\right)\right) = \operatorname{logit}\left(P\left(D_i = j \middle| h_{ij}, y_{ij}, \psi\right)\right) = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}$$

It can be shown that to model a MCAR pattern, one just needs to set ψ_1 and ψ_2 to zero. If only ψ_2 is set to zero, the assumed missing pattern is MAR. If neither ψ_1 or ψ_2 is set to zero, MNAR is assumed. It has been illustrated that it is not appropriate to use the likelihood ratio tests for the individual coefficient ψ_1 or ψ_2 as formal tests for testing the null hypothesis of MAR against the alternative of MNAR (see the discussions from Laird, Little, and Rubin to the original paper by Diggle and Kenward (1994)). Verbeke and Molenberghs (2000) applied the Diggle and Kenward model to longitudinal cases by combining a linear mixed model with a logistic dropout process.

Both design-based and model-based methods work well in terms of reducing biasness if MAR is satisfied (e.g. Horton, and Lipsitz, (2001)). The probability distribution of dropout needs to be specified, which can be difficult due to the limited information that researchers have, if one wants to model the MNAR pattern. There is no uniform way to handle missing data. Some studies suggest that multiple imputation should be used when it is feasible (e.g. Horton, and Kleinman, 2007), while others suggest that using which method used depends on the type of analytical model (e.g. Ibrahim, etc. 2005).

4.3.3. Joint treatment of informative sampling design and dropout

The Most of the longitudinal data available to social scientists is survey data, which means both informative design effects and dropout may be present. A great deal of pioneering researches has been done in this area. For example, Eideh and Nathan (2006, and 2009) proposed an approach that extended Krieger and Pfeffermann's method (1997) to longitudinal data, and combined it with a Diggle and Kenward model.

Let $\mathbf{Y}_i = \begin{pmatrix} Y_{i1} & Y_{i2} & \ldots & Y_{it} \end{pmatrix}$ be a vector of outcomes and $f_p\left(\mathbf{Y}_i|\theta\right)$ denote the population distribution of $\mathbf{Y}_i$. Thus the sample distribution of $\mathbf{Y}_i$ is given by

$$f_s\left(\mathbf{Y}_i|\theta\right) = \frac{E_p\left(\pi_i|\mathbf{Y}_i\right)}{E_p\left(\pi_i\right)} f_p\left(\mathbf{Y}_i|\theta\right)$$

where $E_p\left(\pi_i\right) = \int \ldots \int E_p\left(\pi_i|\mathbf{Y}_i\right) f_p\left(\mathbf{Y}_i|\theta\right) dY_1 \ldots dY_t$.

If we assume the sampling process only happens at the first wave, the probabilities of the sample inclusion only depend on Y_1.

$$f_s\left(\mathbf{Y}_i|\theta\right) = \frac{E_p\left(\pi_i|Y_{i1}\right)}{E_p\left(\pi_i\right)} f_p\left(Y_{i1}|\theta\right) \times \prod_{k=2}^{t} f_p\left(Y_{ik}|\theta\right)$$

Shown by Eideh and Nathan (2009), under the exponential inclusion probability model, the sample distribution of $\mathbf{Y}_i$ follows the same distribution of population $\mathbf{Y}_i$ with the mean shifted by a constant. For example, if $\mathbf{Y}_i$ follows a multivariate normal distribution in population, the sample distribution of $\mathbf{Y}_i$ still is multivariate normal, but its mean differs from the mean of the population distribution, whereas the variance remains the same.

$$f_p\left(\mathbf{Y}_i \middle| \boldsymbol{\beta},\mathbf{V}\right) \sim MVN\left(\mathbf{X}_i'\boldsymbol{\beta},\mathbf{V}_i\right)$$

$$f_s\left(\mathbf{Y}_i \middle| \boldsymbol{\beta},\mathbf{V}_i,a_0\right) \sim MVN\left(\mathbf{X}_i'\boldsymbol{\beta}+a_0v_{11},\mathbf{V}_i\right)$$

where a_0 is a constant, and v_{11} is the first element on the diagonal of $\mathbf{V}_i$.

For an individual who drops out after the second wave but before the last follow-up, under an informative sampling design, the joint likelihood contribution of $\mathbf{y}_i$ is given by,

$$f_s\left(\mathbf{y}_i,d_i \middle| \mathbf{x}_i;\theta,\psi\right)=f_s\left(\mathbf{y}_{i,j-1} \middle| \mathbf{x}_i;\theta\right)\times\prod_{k=2}^{j-1}\left[1-P\left(d_i=k \middle| d_i\geq k,h_{ik},y_{ik},\psi\right)\right]\times P\left(d_i=j \middle| d_i\geq j,h_{ij},y_{ij},\psi\right)$$

where

$$f_s\left(\mathbf{y}_{i,j-1} \middle| \mathbf{x}_i;\theta\right)=\frac{E_p\left(\pi_i \middle| y_{i1}\right)}{E_p\left(\pi_i\right)}f_p\left(y_{i1} \middle| \theta\right)\times\prod_{k=2}^{j-1}f_p\left(y_{ik} \middle| \theta\right) \text{ and}$$

$$P\left(D_i=j \middle| d_i\geq j,h_{ij},y_{ij},\psi\right)=\int P\left(d_i=j \middle| h_{ij},y_{ij},\psi\right)f_p\left(y_{ij} \middle| \theta\right)dy_{ij}$$

The first term refers to the sample distribution of the observed measurements. The second term is a joint probability of dropout before the actual dropout happens. The last term is the model for the actual dropout. Since the dropout model includes the unobserved current value of the response variable, the unobserved value needs to be integrated out given the population distribution of the response variable.

In order to estimate the unknown parameters θ and ψ, a two-stage estimation can be applied. Firstly, a_0 is estimated by regressing $-\log(w_1)$ against y_{i1}, where w_1 is the reciprocal of the probability of inclusion for the first wave. At the second step, the estimated a_0 is substituted into the joint likelihood function. Solving the resulting likelihood function gives the estimates.

4.4 Simulation study

It is still largely unknown which method presented above is superior when both informative sampling and nonignorable missing data are present. A simulation study is conducted to address this question. Three methods are evaluated: panel weighting, multiple imputation and the sample distribution method with the Diggle and Kenward dropout process. These three methods are investigated under MCAR, MAR, and MNAR.

Firstly, a superpopulation with 100,000 individuals and 5 waves is generated according to the following mixed model

$$\mathbf{Y}_i = \mathbf{X}_\mathbf{i}\boldsymbol{\beta} + \mathbf{Z}_i b_i + \boldsymbol{\varepsilon}_i$$

with covariance structure $\operatorname{var}(\mathbf{Y}_i) = \mathbf{V}_i = d\mathbf{J}_t + \sigma^2\mathbf{I}_t + \tau^2\mathbf{H}_t$

where $\boldsymbol{\beta}$ is a vector of fixed effect, d is the variance of random effect b_i, σ^2 is the variance of error terms, and τ^2 is the variance of autoregressive errors. $\mathbf{J}_t$ is a $t \times t$ matrix with all elements equal to 1, $\mathbf{I}_t$ is a $t \times t$ identity matrix, and $\mathbf{H}_t$ is a $t \times t$ matrix with an autoregressive error structure. The error terms across time are correlated to the first order autoregressive structure with a correlation coefficient .8.

Secondly, a sample with 300 individuals is selected by a PPS systemic method. The probabilities of inclusion are proportional to a size variable z which is defined as below

$$z_i = \exp(6 - .2 * y_{1i})$$

A logistic dropout model is used to determine when dropouts happen.

$$\operatorname{logit}\left(P\left(D_i = j \middle| D_i \geq j, h_{ij}, y_{ij}, \psi\right)\right) = \psi_0 + \psi_1 y_{i,j-1} + \psi_2 y_{ij}$$

A MCAR design is obtained by setting $\psi_1 = \psi_2 = 0$. if ψ_2 is set to 0, the dropout process no longer depends on the current measurement of response variable, which implies a MAR. Setting neither ψ_1 or ψ_2 results to a MNAR design. Table 4.1 summarizes all simulation designs.

All simulation designs are generated using SAS 9.2. Each of simulation designs is repeated 500 times. Multiple imputation is accomplished by SAS PROC MI using Expectation Maximization (EM) algorithm with 20 imputations. Each of imputed data sets is analyzed by PROC MIXED, and the final result is obtained by combining all of results using PROC MIANALYZE. Multilevel pseudo likelihood is conducted using GLLAMM in STATA, and the sample distribution estimator is programmed in SAS using PROC IML. The nonlinear optimization of the likelihood function of the sample distribution estimator is carried out by the Newton-Raphson ridge method, and the corresponding Hessian matrix is calculated by the finite differences method using CALL NLPFDD in PROC IML. The empirical Relative Bias (RB), Mean Square Error (MSE), and coverage rates are used to compare the quality of the estimates.

4.5 Results and discussion

In the following the results of simulation are reported. The parameters of interest are regression coefficients and population variance for the linear model. The empirical relative bias, MSE and coverage rates are reported in Table 4.2 to 4.4.

Key findings for each of the tables are as follows,

(1) If the missing pattern is MCAR, both the design-based and the model-based methods performance well. Multiple imputation and the naïve method give very accurate estimates. Except the intercept, all other parameters are very close to the true value, although the coverage rate for β_2 is poor, which is due to slightly positive biases. Multilevel pseudo likelihood produces accurate estimates for the fixed effect. However, the estimated covariance for the random effect is biased.

This is due to that the underlying error structure is not correctly specified. The coverage rates are also low. Consistent with the findings in Chapter 3, when the sampling design is informative, ignoring the sampling design leads to biased estimates. Since the sampling design is correlated to the response variable in an exponential form, the estimated intercept is biased for the naïve method. This is improved by the sample distribution method. The relative bias is reduced substantially from more than 60% to 13%. The coverage rate is improved as well.

(2) Under MAR, some of estimates obtained from the model-based methods and the design-based method are biased. Neither the naïve method nor multiple imputation gives good estimate for the intercept, but the estimates of all other parameters are close to the true values. The multilevel pseudo likelihood method does not work well on covariance parameters, for example the estimated variance for the random effect is biased. The sample distribution method reduces the relative biases and coverage rate of the intercept.

(3) Since multiple imputation and the naïve method assume MAR, one should expect their estimates under MNAR should be worse than those obtained under MAR or MCAR condition. However, this is not observed in Table 4. A possible explanation is that the dropout mechanism used in the simulation only depends on the response variable. The sample distribution method combined with the Diggle and Kenward model improves the estimates of intercept and covariance parameters. Although the multilevel pseudo likelihood method does not work well for either the intercept or the variance of the random effect, it produces accurate estimates for all other parameters.

To summarize, model-based methods perform better compared to the design-based ones in the context of this simulation study. Multiple imputation and the naïve method produce accurate estimates for all parameters except for the intercept. The sample distribution method reduces the bias of the estimated intercept dramatically. The multilevel pseudo likelihood method works well under the most of circumstances, except for the intercept and covariance parameters. It produces biased estimate for the intercept, and less stable estimates for covariance parameters the in terms of MSE.

4.6 Conclusion

In this study, the relative performance of the design-based and the model-based methods for compensating the informative sampling design and dropout has been investigated in a specific statistical setting. The simulation results indicate that both the model-based and the design-based approaches generally work well in the MAR and MNAR settings. In particular, the sample distribution method combined with the Diggle and Kenward model has advantages of correcting the design effect and the nonignorable dropout. The numerical convergence might be an issue for the multilevel pseudo likelihood method. In present study, about 2% of estimation did not converge after 16,000 iterations.

The study has some limitations though. For instance, a rather simple form of informative sampling design is chosen. This leads to the design effect only affecting the intercept. Also, there are no covariates in the dropout model, but they can be easily added in future studies. For example, the dropout can also depend on auxiliary variables or other model covariates. One only needs to change the likelihood function for the logistic model.

CHAPTER 5

AN EMPIRICAL EXAMPLE—ADD HEALTH DATA

5.1 Introduction

The National Longitudinal Study of Adolescent Health (Add Health) is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States during the 1994-1995 school year (Harris, Florey, Tabor, Bearman, Jones, and Udry 2003). It provides a rich set of information on respondents' social, economic, psychological and physical well-being with contextual data on the family, neighborhood, community, school, friendships, peer groups, and romantic relationships. In this chapter, Add Health data is used to test the performance of the methods discussed in Chapter 2-4, and the conclusions from the previous simulation studies are further examined. The analysis strategies used here can be easily adjusted for and applied to other real analysis. In addition, the results reached here may be of interest to researchers who work with Add Health data.

The remainder of the paper is organized as follows. Section 2 introduces the study design of Add Health and the dropout issue. Section 3 outlines the design-based and model-based modeling strategies for the analysis. Section 4 presents the results of the fitted model. The last section summarizes the findings.

5.2 Design and data of Add Health study

Design

From the primary sampling frame collected by Quality Education Data, Inc., 80 high schools were drawn by systematic sampling methods and implicit stratification, which ensured that those selected were representative of US schools with respect to region of country, urbanicity, size, type, and ethnicity. More than 70 percent of the sampled high schools participated. Those who declined to participate were replaced by schools within the same stratum. Each of participating high schools also indentified its feeder schools—schools that have a 7th grade and sent at least five graduates to that high school. One feeder school was selected with a probability proportional to the number of students it contributed to the high school. There were a total of 132 schools in the core study. All students who were listed on a school roster were eligible for being selected selection into the core in-home sample and were stratified by grade and gender. About 17 students were randomly chosen from each stratum and approximately 200 from each of the 80 pairs of schools (school and its feeder school). In summary, 12,105 adolescents were interviewed in the core sample. Add Health also oversampled the minority populations such as African Americans from highly-educated families, Chinese, Cuban, Puerto Rican, the disabled, and twins.

The Add Health study was designed as a cluster sample in which the clusters were sampled with unequal probabilities. Chart 5.1 shows the basic sampling strategy of Add Health study. The total sample size of Wave 1 was 20,745, which included the core sample, saturation samples, the disabled sample, and the ethnic sample. The probability of inclusion varies across samples.

Non-response

Non-response occurred at each stage of the sampling. For instance, 52 out of the 80 initially selected schools agreed to cooperate. In four cases, researchers failed to recruit a feeder school to take part in the study.

It has been reported that the Add Health follow-up nonresponse was associated with many factors that had influence on four types of recruitment outcomes: contactability, unwillingness, inability, and participation Those factors include: the degree of urbanization, neighborhood safety for the local social environment, household measures (e.g. size, income, the number of years at the current address, the type of residential structure, residents' ages and the social inter-relationships), parental characteristics (e.g., the education level, employment status, involvement in relevant organizations), respondent's characteristics (e.g., age, race, gender, religion, school experience, and substance use), and the experience from prior waves (e.g. bored, embarrassed) (Kalsbeek, Yang and Agans, 2002).

5.3 Analysis strategies

To compare the relative performance of the design-based and the model-based approaches in addressing the statistical implications of the sample design and dropout, two cross-sectional and one longitudinal models are estimated. The cross-sectional models only use the first wave data; while the longitudinal model uses three waves data. For each of models, both the design-based and model-based methods are used. For design-based approaches, weights are included into a model, and the adjustment according to sampling design features is implemented. For the model-based approach, the naïve method with the robust standard error, the sample distribution method and multiple imputation (for the longitudinal model) are implemented. Table 5.1 summarizes the modeling strategies for each of models.

5.3.1 Design-based approaches

To compensate for the design effect of Add Health, a complex weighting process has been implemented to generate grand sample weights. Each of the in-home samples was weighted in four main steps. First, a preliminary school weight (which was the reciprocal of the school's probability of inclusion) was calculated to compensate for the unequal probability of inclusion of schools. Second, this weight (W1) was adjusted for school ineligibility and nonresponse among eligible schools. In addition, the adjusted school weight (W2) was adjusted again to match population totals according to post stratification. Third, an initial individual-level conditional within-school weight (W3) was calculated to compensate for differences in individual selection probabilities across schools and across grades and genders within a school. The final weight (W4) was generated to adjust for both school-level and individual-level nonresponse (Tourangeau, and Shin, 1999).

As mentioned by many researchers (e.g. Pfeffermann, et al., 1998, 24), single-level weights may not carry adequate information to correct for the higher level unequal probabilities of inclusion. To incorporate the multiple stage sampling design, multilevel weights are needed.

For the Add Health study, the level 2 school weights (W2) were constructed to compensate for school-level unequal selection probabilities and nonresponse. Level 1 individual weights have to be calculated using the final weights (W4) divided by school weights (W2) (Chantala, Blanchette and Suchindran, 2006).

Longitudinal weights were also created using the sample response rate to compensate for dropouts. For example, the Wave II weight was calculated as Wave I weight over the response rate for that sample in Wave II, if the sample to be used for longitudinal analysis consists of the Wave II respondents.

5.3.2 Model-based approaches

Sample distribution estimators

For the single-level cross-sectional model, a test for the informativeness is conducted to determine the appropriate form of the conditional expectation under an exponential approximation. Firstly, unknown parameters in $E_p(\pi_i \mid y_i)$ are estimated using the quadratic form

$$E_s(\pi_i \mid y_i, x_i) = \exp\left(a_0 + a_1 y_i + a_2 y_i^2\right)$$

Then, a t test is conducted for the quadratic term by regressing w_i against $\hat{\varepsilon}_i^2$ to check if the coefficient equals to zero, where $\hat{\varepsilon}_i^2$ is the squared ordinary least square residual term from regressing independent variables against the response variable. If so, the quadratic term will be dropped.

For the multilevel cross-sectional model, a first order exponential approximation is used for both levels. Similar as in Chapter 3, $\bar{y}_i$ (school mean) is used as a measure of μ_i. Then the estimated b_1 and d_1 are substituted to the full sample likelihood function, and solving the likelihood function gives the estimates for the unknown parameters.

For the longitudinal model, since the sampling process only happened at the first wave, the probabilities of the sample inclusion only depend on Y_1. Under the exponential inclusion probability model, the sample distribution of $\mathbf{Y}_i$ follows the same distribution of population $\mathbf{Y}_i$ with the mean shifted by a constant.

$$f_p\left(\mathbf{Y}_i \middle| \boldsymbol{\beta}, \mathbf{V}\right) \sim MVN\left(\mathbf{X}_i'\boldsymbol{\beta}, \mathbf{V}_i\right)$$

$$f_s\left(\mathbf{Y}_i \middle| \boldsymbol{\beta}, \mathbf{V}_i, a_0\right) \sim MVN\left(\mathbf{X}_i'\boldsymbol{\beta} + a_0 v_{11}, \mathbf{V}_i\right)$$

where a_0 is a constant, and v_{11} is the first element on the diagonal of $\mathbf{V}_i$.

To model the missing patterns, a Diggle and Kenward model is proposed for all three missing patterns. To simplify the analysis, the dropout process is assumed to be dependent only on the response variable.

Naïve method

For comparative purpose, a naïve method is also used. For the single-level cross-sectional model, OLS is used, and robust standard errors are estimated to compensate for the clustering effect. For the multilevel cross-sectional model, a school-specified random effect model is estimated, and the cluster robust standard errors are also reported.

For the longitudinal model, an individual-specified random effect is estimated. To address the unit dropout issue and the correlation among repeated measures across waves, a separate longitudinal model with AR(1) structure is estimated. Then multiple imputation with 20 replicates using EM algorithm is implemented to compensate the unit dropout.

5.4 Results and discussion

The response variable in this study is the adolescent delinquency behaviors. In accordance with the literature on adolescent delinquency (Hagan and Foster 2003; Hannon 2003; Haynie 2001; 2003), a delinquency scale containing 12 standard items is developed. These items include: stealing amounts larger or smaller than $50, breaking and entering, drug selling, serious physical fighting that resulted in injuries needing medical treatment, use of weapons to get something from someone, involvement of physical fighting between groups, shooting or stabbing someone, deliberately damaging property, and pulling a knife or gun on someone. For the cross-sectional model, Wave I delinquency score is the response variable. For the longitudinal model, the response is a vector that contains measurements of delinquency score across waves.

The predictors are categorized into two groups. The first group consists of the structural and demographic variables, such as age at baseline and race (White, Hispanic, Black, and Other). The second group consists of the family and school process variables. These include living with two biological parents (0-1), parental unemployment (dummy where 1 indicates whether one or both parents were unemployed at baseline, education level of the adult interviewed at baseline (less than high school, high school, at least some college), daily family meals (dummy indicating whether at least one of the respondent's parents was in the room with he/she while he/she ate his/her evening meal at least six of the past seven days, and repeating a grade (0-1).

Table 5.2 presents basic statistics of the variables in this study. For simplicity's sake, all predictors in this study are from Wave I, which means only missing response values need to be imputed.

Table 5.3 reports the regression results of cross-sectional models. For the single level cross-sectional models, different modeling strategies do not lead to much discrepancy. For example, all estimates from the pseudo likelihood method, the sample distribution method and the naïve OLS are in the same direction. Except the intercept and the variance, all estimates are very close. Since the sample distribution method uses a second order exponential approximation for the conditional probabilities of the sampling inclusion, both the intercept and the variance are different from the OLS estimates.

For the design-based cross-sectional multilevel model, both the school level and the individual level weights are used to compensate for the unequal selection probabilities at both school level and individual level. The sample distribution method uses the school mean as a measure of μ_i, then the estimated b_1 and d_1 are substituted to the full sample likelihood function. Consistent to the results in Chapter 3, the only difference between the estimates from the sample distribution method and those from the naïve multilevel model is the intercept. The estimated standard errors are also slightly different, but do not lead to different inferences. Most of the estimated coefficients from the weighted multilevel model are apart from those estimated by either the sample distribution method or the naïve multilevel model. This is due to the effects of the school level sampling weights. The model-based methods produce very close estimates in both the single level and the multilevel models. The design-based estimates are not consistent across models.

Table 5.4 summarizes all the design-based and the model-based estimates for the longitudinal models. The weighted multilevel model uses the school level and the individual level weights to adjust for the sampling design and dropouts. Suggested by the Add Health researchers, the grand panel weight (GSWGT3) should be used if the analysis involves respondents who were interviewed at all three waves. No weighting suggestions are given for analysis that involves all available respondents at all three waves. However, using the grand panel weight (GSWGT3) as the individual level weights leads to numerical problems in this study-- GLLAMM in STATA did not converge after 24 hours iterations.

As an alternative, a panel weight is used for this study. For the individuals in wave 1, the grand weight (GSWGT1) is used, while for those in the later waves, the panel weights that only compensate for dropouts are used. For example, the panel weight for the second wave is defined as the grand weight for the second wave over the grand weight for the first wave.

The first two columns in Table 4 present the results from the weighted and the naïve multilevel models. Since neither of these two models considers the repeated measure of individuals, they are comparable in terms of the model specification. It can be seen that none of estimated coefficients is very close, but most of them are in the same direction. In addition, the inferences based on those estimates are the same.

To take into account the correlation among repeated measures and the dropout process, two model-based approaches are implemented: one is the sample distribution method with the Diggle and Kenward model, and the other one is multiple imputation. The sample distribution method does not give any estimates that are consistent to other methods, no matter which missing pattern is assumed. In fact, the estimation is sensitive to the initial values for MNAR. As mentioned by Molenberghs and etc (2007), the Diggle and Kenward model involves a computationally intensive numerical integration. When it is combined with the sample likelihood function, the likelihood surface tends to be rather flat. Therefore, the estimation process is relatively easy to achieve a local maximum. As presented here, the sample distribution method with the Diggle and Kenward model does not work well. In a word, the sample distribution method combined with a logistic dropout is not easy to use in the real data analysis.

The results from multiple imputation are consistent to those obtained from the naïve and the weighted method. This indicates that the multiple imputation might not be a bad choice for this type of analysis.

5.5 Conclusion

In this chapter, the design-based and the model-based methods to compensate for the sampling design and the dropout are implemented using Add Health study as an example. The results from both the design-based and the model-based methods are discussed. Although guidelines for how to pick up the correct weights have been given by Add Health researchers, researchers should be aware of the effect of the sampling weights on the model estimation.

Previous chapters have discussed the theory and practice of the sample distribution method. In this chapter, the sample distribution method is implemented in the real data analysis along with the test of the informativeness. In fact, the correlation between the response variable and the sampling design is not strong in this analysis. Based on the evidences in Chapter 2-4, the estimates from the sample distribution method or the naïve method should be trusted for the cross-sectional models.

The sample distribution method combined with the Diggle and Kenward model has been showed to have some advantages of correcting for the design effect and nonignorable dropout in Chapter 4, However, it does not work well in this analysis. For the longitudinal models, either multilevel weighted method or multiple imputation can be used, since the estimates obtained from these two methods are close and the inferences are the same.

APPENDIX

Table 2.1 Simulation Design

Design		1	2	3	4
Informativeness	Informative	√		√	
	Non informative		√		√
Estimators	Design-based: PML	√	√	√	√
	Sample Dist: W1s	√	√	√	√
	Sample Dist: W2s	√	√	√	√
	Naïve: OLS/ML	√	√	√	√
Model type	Linear model	√	√		
	Non-linear model			√	√

Table 2.2 Relative Biases and Mean Square Errors (MSE) of Linear Model:

Exponential Sampling with 1st Order Correlation.

Estimator	*Parameter*	*Relative bias(MSE)*	*Coverage*
OLS	β_0	-0.424(0.252)	0.000
	β_1	-0.009(0.146)	0.890
	β_2	-0.021(0.023)	0.430
	β_3	0.016(0.015)	0.826
	σ^2	-0.041(0.399)	0.518
PML	β_0	-0.049(0.871)	0.718
	β_1	-0.053(0.707)	0.872
	β_2	-0.044(0.118)	0.646
	β_3	0.041(0.06)	0.776
	σ^2	-0.135(1.718)	0.490
W1s	β_0	-0.036(0.271)	0.610
	β_1	-0.009(0.146)	0.890
	β_2	-0.021(0.023)	0.430
	β_3	0.016(0.015)	0.826
	σ^2	-0.013(0.051)	0.750
W2s	β_0	-0.341(0.29)	0.000
	β_1	-0.001(0.177)	0.918
	β_2	-0.008(0.030)	0.860
	β_3	0.067(0.017)	0.368
	σ^2	-0.014(0.505)	0.846

Table 2.3 Relative Biases and Mean Square Errors (MSE) of Linear Model: Exponential Sampling with 2nd Order Correlation.

Estimator	*Parameter*	*Relative bias(MSE)*	*Coverage*
OLS	β_0	0.206(0.656)	0.088
	β_1	-0.150(0.136)	0.702
	β_2	-0.137(0.025)	0.000
	β_3	-0.137(0.036)	0.384
	σ^2	-0.140(0.355)	0.000
PML	β_0	0.106(1.988)	0.774
	β_1	-0.026(0.849)	0.886
	β_2	-0.028(0.134)	0.816
	β_3	-0.129(0.123)	0.772
	σ^2	-0.195(2.350)	0.552
W1s	β_0	-0.094(1.458)	0.860
	β_1	0.020(0.501)	0.898
	β_2	0.004(0.097)	0.774
	β_3	0.009(0.133)	0.902
	σ^2	-0.021(1.49)	0.832
W2s	β_0	0.188(1.131)	0.770
	β_1	-0.108(0.441)	0.910
	β_2	-0.121(0.081)	0.106
	β_3	-0.131(0.116)	0.856
	σ^2	-0.143(1.144)	0.390

Table 2.4 Relative Biases and Mean Square Errors (MSE) of Linear Model:

Non-informative Sampling.

Estimator	*Parameter*	*Relative bias(MSE)*	*Coverage*
OLS	β_0	-0.003(0.338)	0.902
	β_1	0.003(0.146)	0.900
	β_2	0.000(0.023)	0.908
	β_3	0.003(0.020)	0.902
	σ^2	-0.003(0.412)	0.870
PLM	β_0	0.000(0.595)	0.856
	β_1	0.000(0.379)	0.868
	β_2	0.001(0.059)	0.872
	β_3	-0.002(0.04)	0.880
	σ^2	-0.028(1.015)	0.796
W1s	β_0	-0.042(0.339)	0.674
	β_1	0.005(0.146)	0.900
	β_2	0.002(0.023)	0.902
	β_3	0.005(0.020)	0.890
	σ^2	0.002(0.013))	0.874
W2s	β_0	-0.003(0.34)	0.908
	β_1	0.003(0.147)	0.896
	β_2	0.000(0.023)	0.908
	β_3	0.003(0.02)	0.888
	σ^2	-0.003(0.415)	0.870

Table 2.5. Relative Biases and Mean Square Errors (MSE) of Logistic Model:

Exponential Sampling with 1st Order Correlation

Estimator	*Parameter*	*Relative bias(MSE)*	*Coverage*
OLS	β_0	-0.243(0.439)	0.832
	β_1	0.037(0.250)	0.914
	β_2	0.026(0.140)	0.882
	β_3	0.028(0.042)	0.89
PLM	β_0	0.026(0.738)	0.856
	β_1	0.153(0.583)	0.858
	β_2	0.139(0.318)	0.73
	β_3	0.150(0.091)	0.728
W1s	β_0	0.003(0.434)	0.888
	β_1	0.028(0.246)	0.926
	β_2	0.017(0.125)	0.89
	β_3	0.019(0.040)	0.898
W2s	β_0	-0.224(0.440)	0.844
	β_1	0.039(0.251)	0.924
	β_2	0.027(0.140)	0.884
	β_3	0.032(0.043)	0.888

Table 2.6 Relative Biases and Mean Square Errors (MSE) of Logistic Model: Non-informative Sampling.

Estimator	*Parameter*	*Relative bias(MSE)*	*Coverage*
OLS	β_0	0.045(0.468)	0.91
	β_1	0.046(0.267)	0.884
	β_2	0.030(0.150)	0.874
	β_3	0.025(0.045)	0.862
PLM	β_0	0.188(0.787)	0.858
	β_1	0.176(0.610)	0.864
	β_2	0.144(0.337)	0.728
	β_3	0.129(0.097)	0.796
W1s	β_0	0.037(0.463)	0.91
	β_1	0.035(0.262)	0.886
	β_2	0.02(0.131)	0.87
	β_3	0.015(0.042)	0.866
W2s	β_0	0.048(0.471)	0.91
	β_1	0.047(0.269)	0.894
	β_2	0.030(0.150)	0.876
	β_3	0.024(0.045)	0.864

Table 2.7. Test for the Informativeness

Model	*Correlation with the response*	*Residual*	*N*	*Proportion of rejection*
Linear	1st order	OLS	500	0.969
	2nd order	OLS	500	0. 840
	Non-informative	OLS	500	0.018
Logistic	1st order	deviance	500	0.660
	1st order	Pearson	500	0.138
	Non-informative	deviance	500	0.282
	Non-informative	Pearson	500	0.056

Table 2.8 Relative Biases and Mean Square Errors (MSE) of Linear Model:

Estimator	*Parameter*	*Relative bias(MSE)*	*Coverage*
W1s	β_0	-0.054(0.754)	0.586
	β_1	-0.029(0.155)	0.850
	β_2	-0.015(0.030)	0.654
	β_3	-0.014(0.041)	0.788
	σ^2	-0.018(0.464)	0.726
W2s	β_0	0.188(0.679)	0.158
	β_1	-0.120(0.143)	0.780
	β_2	-0.107(0.026)	0.000
	β_3	-0.136(0.037)	0.428
	σ^2	-0.109(0.388)	0.010

Table 3.1 Simulation Design

Design			1	2	3	4
Informativeness	Sampling stage	1st	√	√		
		2nd	√		√	
Methods	Design-based :MPML		√	√	√	√
	Sample Distribution		√	√	√	√
	ML with robust stderr		√	√	√	√

Table 3.2 Relative Biases, Mean Square Errors, and Coverage Rates of Three Estimation Methods for Design 1: Both Stages are Informative.

Methods	*Parameter*	*Relative bias(MSE)*	*Coverage*
Naïve Multilevel	β_0	-0.399(0.280)	0.000
	β_1	-0.016(0.150)	0.948
	β_2	-0.012(0.026)	0.830
	β_3	-0.007(0.015)	0.938
	σ_u^2	-0.160(0.348)	0.666
	σ_e^2	-0.015(0.432)	0.918
Weighted Multilevel	β_0	-0.607(1.058)	0.151
	β_1	-0.018(0.525)	0.896
	β_2	0.035(0.093)	0.816
	β_3	0.037(0.053)	0.888
	σ_u^2	-0.167(0.751)	0.495
	σ_e^2	-0.055(1.432)	0.753
Sample Distribution	β_0	0.252(0.303)	0.282
	β_1	-0.016(0.150)	0.992
	β_2	-0.012(0.026)	0.952
	β_3	-0.007(0.015)	0.988
	σ_u^2	-0.160(0.348)	0.878
	σ_e^2	-0.015(0.432)	0.984

Table 3.3 Relative Biases, Mean Square Errors, and Coverage Rates of Three Estimation Methods for Design 3: 1st Stage is Informative.

Methods	*Parameter*	*Relative bias(MSE)*	*Coverage*
Naïve Multilevel	β_0	0.373(0.637)	0.35
	β_1	0.010(0.28)	0.922
	β_2	0.000(0.045)	0.938
	β_3	-0.006(0.032)	0.946
	σ_u^2	-0.034(0.879)	0.886
	σ_e^2	-0.004(0.762)	0.956
Weighted Multilevel	β_0	0.284(0.878)	0.618
	β_1	-0.011(0.387)	0.938
	β_2	-0.002(0.064)	0.928
	β_3	-0.002(0.044)	0.920
	σ_u^2	-0.101(0.651)	0.532
	σ_e^2	-0.074(0.971)	0.658
Sample Distribution	β_0	0.146(0.744)	0.854
	β_1	0.010(0.28)	0.936
	β_2	0.000(0.045)	0.944
	β_3	-0.006(0.032)	0.948
	σ_u^2	-0.034(0.879)	0.886
	σ_e^2	-0.004(0.762)	0.956

Table 3.4 Relative Biases, Mean Square Errors, and Coverage Rates of Three Estimation Methods for Design 3: 2nd Stage is Informative.

Methods	*Parameter*	*Relative bias(MSE)*	*Coverage*
Naïve Multilevel	β_0	0.373(0.637)	0.350
	β_1	0.010(0.28)	0.922
	β_2	0.000(0.045)	0.938
	β_3	-0.006(0.032)	0.946
	σ_u^2	-0.034(0.879)	0.886
	σ_e^2	-0.004(0.762)	0.956
Weighted Multilevel	β_0	-0.902(0.986)	0.016
	β_1	0.057(0.512)	0.910
	β_2	0.033(0.08)	0.788
	β_3	0.044(0.048)	0.864
	σ_u^2	0.133(0.789)	0.784
	σ_e^2	-0.044(1.411)	0.790
Sample Distribution	β_0	0.490(0.646)	0.148
	β_1	0.010(0.28)	0.936
	β_2	0.000(0.045)	0.944
	β_3	-0.006(0.032)	0.948
	σ_u^2	-0.034(0.879)	0.886
	σ_e^2	-0.004(0.762)	0.956

Table 3.5 Relative Biases, Mean Square Errors, and Coverage Rates of Three Estimation Methods for Design 4: Neither Stage is Informative.

Methods	*Parameter*	*Relative bias(MSE)*	*Coverage*
Naïve Multilevel	β_0	-0.002(0.671)	0.926
	β_1	0.004(0.266)	0.948
	β_2	0.000(0.046)	0.940
	β_3	0.001(0.034)	0.904
	σ_u^2	-0.058(0.865)	0.886
	σ_e^2	-0.004(0.762)	0.956
Weighted Multilevel	β_0	0.006(0.741)	0.904
	β_1	-0.006(0.291)	0.958
	β_2	-0.002(0.051)	0.934
	β_3	-0.002(0.036)	0.916
	σ_u^2	0.129(0.625)	0.780
	σ_e^2	-0.075(0.84)	0.574
Sample Distribution	β_0	-0.002(0.671)	0.950
	β_1	0.004(0.266)	0.960
	β_2	0.000(0.046)	0.952
	β_3	0.001(0.034)	0.934
	σ_u^2	-0.057(0.865)	0.886
	σ_e^2	-0.004(0.776)	0.942

Table 4.1 Simulation design

Design		1	2	3
Informative	1st wave	√	√	√
Design-based Method	Panel weighting	√	√	√
Model-based Method	Naïve model	√	√	√
	Multiple Imputation	√	√	√
	Sample distribution	√	√	√
Missing Mechanism	MCAR	√		
	MAR		√	
	MNAR			√

Table 4.2 1st Stage Informative Sampling with MCAR Dropout

Method	*Parameter*		*Relative bias(MSE)*	*Coverage*
Multiple Imputation	Fixed Effect	β_0	-0.638(0.39)	0.000
		β_1	0.062(0.213)	0.916
		β_2	0.057(0.029)	0.030
		β_3	0.056(0.017)	0.608
	Covariance	ρ	-0.001(0.113)	0.942
		σ^2	0.073(0.987)	0.880
		d	0.013(0.89)	0.940
		τ^2	-0.020(0.428)	0.934
Naïve Multilevel	Fixed Effect	β_0	-0.639(0.387)	0.000
		β_1	0.062(0.213)	0.908
		β_2	0.057(0.029)	0.030
		β_3	0.056(0.017)	0.614
	Covariance	ρ	0.002(0.115)	0.936
		σ^2	0.037(1.036)	0.912
		d	0.066(0.94)	0.948
		τ^2	0.040(0.453)	0.936
Weighted Multilevel	Fixed Effect	β_0	-0.603(1.056)	0.150
		β_1	-0.040(0.527)	0.908
		β_2	0.035(0.09)	0.825
		β_3	0.035(0.054)	0.883
	Covariance	σ^2	-0.052(1.418)	0.753
		d	-0.160(0.798)	0.494
Sample Distribution Method with a Logistic Dropout	Fixed Effect	β_0	0.133(0.407)	0.756
		β_1	0.062(0.213)	0.908
		β_2	0.057(0.029)	0.030
		β_3	0.056(0.017)	0.614
	Covariance	ρ	0.002(0.115)	0.970
		σ^2	0.037(1.036)	0.926

Method	*Parameter*		*Relative bias(MSE)*	*Coverage*
		d	0.147(0.94)	0.976
		τ^2	-0.033(0.453)	0.920
	Dropout	ψ_0	-0.001(0.217)	0.970

Table 4.3 1st Stage Informative Sampling with MAR Dropout

Method	*Parameter*		*Relative bias(MSE)*	*Coverage*
Multiple Imputation	Fixed Effect	β_0	-0.664(0.403)	0.000
		β_1	0.061(0.228)	0.913
		β_2	0.059(0.031)	0.032
		β_3	0.058(0.019)	0.638
	Covariance	ρ	-0.008(0.119)	0.962
		σ^2	0.156(1.031)	0.756
		d	-0.094(0.878)	0.931
		τ^2	-0.177(0.393)	0.782
Naïve Multilevel	Fixed Effect	β_0	-0.666(0.402)	0.000
		β_1	0.061(0.226)	0.914
		β_2	0.060(0.03)	0.026
		β_3	0.059(0.019)	0.616
	Covariance	ρ	0.007(0.126)	0.908
		σ^2	0.042(1.155)	0.890
		d	0.072(1.026)	0.924
		τ^2	0.013(0.47)	0.938
Weighted Multilevel	Fixed Effect	β_0	0.284(0.878)	0.618
		β_1	-0.011(0.387)	0.938
		β_2	-0.002(0.064)	0.928
		β_3	-0.002(0.044)	0.920
	Covariance	σ^2	-0.074(0.971)	0.658
		d	-0.101(0.651)	0.532
Sample Distribution Method with a Logistic Dropout	Fixed Effect	β_0	0.105(0.419)	0.842
		β_1	0.061(0.226)	0.916
		β_2	0.060(0.03)	0.028
		β_3	0.059(0.019)	0.616
	Covariance	ρ	0.007(0.126)	0.974
		σ^2	0.042(1.155)	0.916
		d	0.154(1.026)	0.974

Method	*Parameter*		*Relative bias(MSE)*	*Coverage*
		τ^2	-0.058(0.47)	0.894
	Dropout	ψ_0	0.007(0.600)	0.956
		ψ_1	0.011(0.016)	0.952

Table 4.4 1st Stage Informative Sampling with MNAR Dropout

Method	*Parameter*		*Relative bias(MSE)*	*Coverage*
Multiple Imputation	Fixed Effect	β_0	-0.605(0.408)	0.000
		β_1	0.053(0.223)	0.907
		β_2	0.048(0.032)	0.121
		β_3	0.047(0.019)	0.740
	Covariance	ρ	-0.011(0.125)	0.960
		σ^2	0.126(1.035)	0.802
		d	-0.060(0.909)	0.915
		τ^2	-0.189(0.390)	0.750
Naïve Multilevel	Fixed Effect	β_0	-0.610(0.406)	0.000
		β_1	0.053(0.22)	0.912
		β_2	0.049(0.031)	0.086
		β_3	0.048(0.019)	0.700
	Covariance	ρ	-0.002(0.13)	0.918
		σ^2	0.021(1.162)	0.908
		d	0.103(1.059)	0.938
		τ^2	-0.020(0.462)	0.918
Weighted Multilevel	Fixed Effect	β_0	-0.902(0.986)	0.016
		β_1	0.057(0.512)	0.910
		β_2	0.033(0.080)	0.788
		β_3	0.044(0.048)	0.864
	Covariance	σ^2	-0.044(1.411)	0.790
		d	0.133(0.789)	0.784
Sample Distribution Method with a Logistic Dropout	Fixed Effect	β_0	0.105(0.423)	0.852
		β_1	0.063(0.220)	0.908
		β_2	0.060(0.031)	0.026
		β_3	0.059(0.019)	0.604
	Covariance	ρ	0.005(0.129)	0.974
		σ^2	0.036(1.13)	0.928

Method	*Parameter*		*Relative bias(MSE)*	*Coverage*
		d	0.152(1.025)	0.984
		τ^2	-0.070(0.468)	0.884
	Dropout	ψ_0	-0.035(1.16)	0.926
		ψ_1	-0.004(0.018)	0.944
		ψ_2	-0.070(0.023)	0.916

Table 5.1 List of Modeling strategies

Model	Design-based	Model-based method	
	Weights used	Sample distribution	Naïve
Cross-sectional Single level	Cross sectional weights	Exponential approximation	
Cross-sectional Multilevel	2-level cross sectional weights	2-level exponential approximation	robust standard error
Longitudinal	2-level longitudinal weights	Diggle and Kenward dropout	Multiple imputation

Table 5.2 Variable description, means, and standard deviations

Variable name	Description	Mean	SD
Serious and violent delinquency			
Wave I	Serious Delinquency Scale, Wave 1	1.681	3.097
Wave II	Serious Delinquency Scale, Wave 2	1.296	2.759
Wave III	Serious Delinquency Scale, Wave 3	0.763	1.846
Structural/Demographic			
Age/ethnicity			
Age	age at time of interview at Wave I	15.413	1.731
Black	race reported as black at Wave I	0.228	0.419
Hispanic	race reported as Hispanic at Wave I	0.126	0.331
Asian	race reported as Asian at Wave I	0.043	0.202
Family SES			
Parent jobless	Parent Unemployed at Wave I	0.047	0.212
High school	Parent has High School education only Wave I	0.029	0.168
> High school	Parent has education beyond high school	0.566	0.496
Family & School Process			
Daily family meals	Eating meals with parent 6 days / week at Wave I	0.482	0.500
2 biological parents	Living with both parents at Wave I	0.523	0.499
Repeated a grade	Having repeated grade by Wave I	0.210	0.407

Table 5.3 The estimated regression coefficients and standard errors for cross-sectional models.

	Single level			Multilevel		
Parameter	Design-based	Model-based		Design-based	Model-based	
	PML	Sample Distribution Method	Naïve OLS	Weighted Multilevel Model	Sample Distribution Method	Naïve Multilevel
Intercept	-7.762(2.305)***	-8.156(1.884)***	-8.162(2.033)***	3.117(0.506)***	-7.973(2.888)**	-7.841(2.019)***
Age	1.391(0.300)***	1.434(0.248)***	1.434(0.266)***	0.063(0.049)*	1.396(0.373)***	1.395(0.265)***
Age^2	-0.048(0.010)***	-0.049(0.008)***	-0.049(0.009)***	-0.006(0.001)***	-0.048(0.012)***	-0.048(0.009)***
Black	0.045(0.104)	-0.015(0.072)	-0.015(0.092)	0.128(0.105)	0.000(0.114)	0.000(0.083)
Hispanic	0.554(0.137)***	0.577(0.103)***	0.577(0.106)***	0.736(0.328)**	0.531(0.135)***	0.531(0.107)***
Asian	0.319(0.219)*	0.077(0.128)	0.077(0.099)	0.672(0.585)	0.086(0.210)	0.086(0.116)
Repeated grade	0.823(0.112)***	0.767(0.082)***	0.767(0.080)***	0.948(0.141)***	0.786(0.101)***	0.787(0.079)***
PVT 90-110	-0.064(0.118)	-0.057(0.087)	-0.057(0.099)	-0.234(0.222)	-0.069(0.123)	-0.069(0.096)
PVT 110+	0.031(0.078)	0.065(0.061)	0.065(0.063)	-0.076(0.129)	0.049(0.089)	0.049(0.061)
Religious	-0.35(0.071)***	-0.409(0.054)***	-0.409(0.059)***	-0.194(0.080)**	-0.383(0.081)***	-0.383(0.059)***
2-bio parents	-0.46(0.074)***	-0.409(0.056)***	-0.409(0.053)***	-0.241(0.155)*	-0.373(0.080)***	-0.373(0.052)***
Unemployed	0.553(0.206)**	0.382(0.158)**	0.382(0.178)**	0.137(0.172)	0.378(0.178)**	0.378(0.181)**
High School	-0.543(0.191)**	-0.444(0.156)**	-0.444(0.206)**	-0.395(0.294)*	-0.391(0.241)*	-0.391(0.185)**

Some college+	-0.038(0.078)	-0.002(0.06)	-0.002(0.06)	0.051(0.094)	-0.027(0.085)	-0.027(0.060)
Family Meal	-0.624(0.076)***	-0.535(0.057)***	-0.535(0.063)***	-0.733(0.177)***	-0.537(0.080)***	-0.537(0.062)***
σ_e^2	9.311(0.401)***	1.742(0.017)***	9.225(0.115)***	8.940(1.546)***	9.120(0.162)***	9.120(0.115)***
σ_u^2				0.170(0.008)***	0.110(0.038)**	0.110(0.027)***

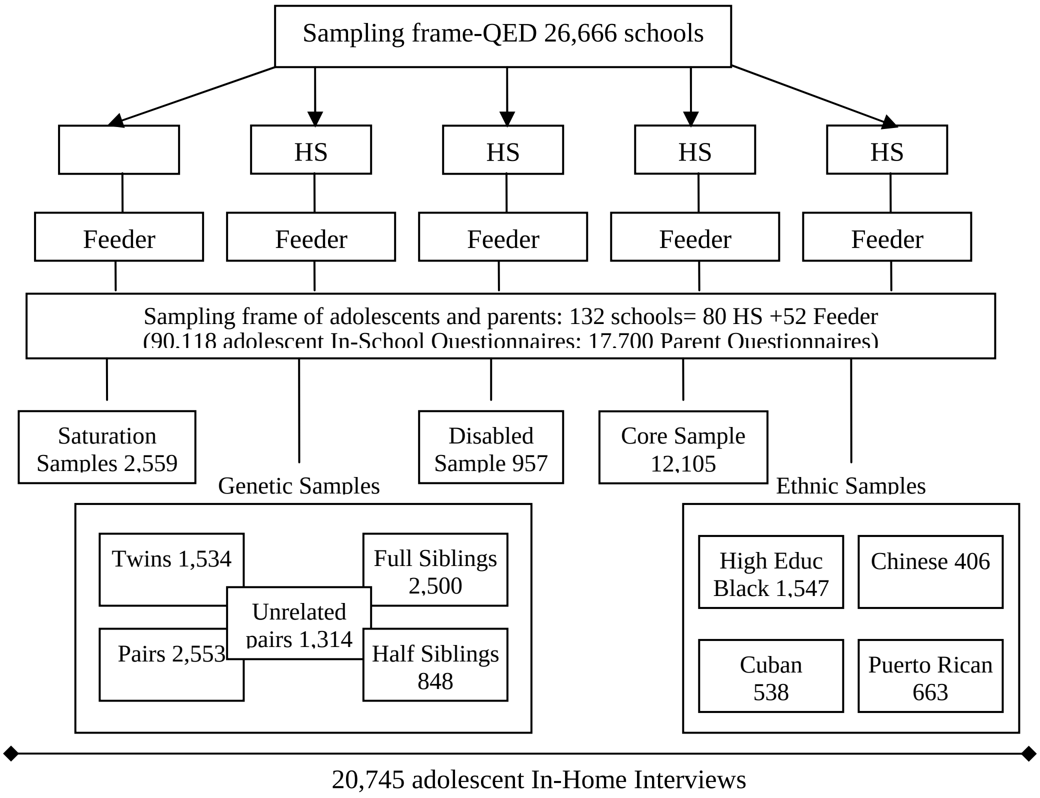
Table 5.4 The estimated coefficients and standard errors for longitudinal models

	Design-based	Model-based				
Parameter	Weighted Multilevel Model*	Naïve Multilevel Model	Sample Distribution Method with Logistic Dropout			Multiple Imputation
			MCAR	MAR	MNAR	
Intercept	-11.399(2.191)***	-2.194(1.248)**	-20.592(17.308)	-20.596(16.898)	-27.687(17.353)*	-2.153(1.388)
Age	1.847(0.237)***	0.590(0.162)***	3.053(2.316)*	3.053(2.264)*	3.965(2.323)**	0.589(0.182)**
Age^2	-0.062(0.008)***	-0.022(0.005)***	-0.107(0.077)*	-0.107(0.075)*	-0.136(0.077)**	-0.022(0.006)***
Black	0.040(0.075)	0.049(0.050)	-0.627(0.654)	-0.627(0.653)	-0.51(0.661)	0.034(0.054)
Hispanic	0.575(0.086)***	0.433(0.058)***	0.263(0.732)	0.263(0.728)	0.308(0.742)	0.454(0.062)***
Asian	0.391(0.154)**	0.045(0.090)	5.905(3.11)**	5.905(3.11)**	6.52(3.103)**	0.028(0.099)
Repeated grade	0.815(0.065)***	0.532(0.042)***	2.161(0.639)***	2.161(0.638)***	2.107(0.646)***	0.532(0.047)***
PVT 90-110	-0.142(0.08)**	-0.081(0.051)*	-1.553(0.814)**	-1.553(0.81)**	-1.289(0.828)*	-0.097(0.058)*
PVT 110+	0.043(0.058)	-0.015(0.037)	-1.253(0.709)**	-1.253(0.709)**	-1.049(0.722)*	-0.021(0.041)
Religious	-0.396(0.053)***	-0.32(0.034)***	-0.206(0.567)	-0.206(0.565)	-0.208(0.576)	-0.322(0.038)***
2-bio parents	-0.461(0.053)***	-0.27(0.033)***	0.023(0.652)	0.023(0.649)	0.411(0.657)	-0.287(0.036)***
Unemployed	0.480(0.116)***	0.233(0.074)***	1.154(1.100)	1.154(1.100)	1.328(1.113)	0.253(0.083)**
High School	-0.510(0.154)***	-0.262(0.101)**	-0.765(1.136)	-0.765(1.132)	-1.189(1.166)	-0.305(0.111)**

Some college+	0.003(0.054)	0.038(0.036)	-0.852(0.6)*	-0.852(0.598)*	-0.942(0.608)*	0.051(0.04)
Family Meal	-0.621(0.052)***	-0.344(0.033)***	-0.995(0.538)**	-0.995(0.539)**	-0.867(0.548)*	-0.351(0.038)***
d	0.163(0.687)	0.095(0.017)***	4.397(1.633)**	4.397(1.646)**	5.083(0.791)***	0.046(0.011)***
τ^2			6.940(14.169)	5.243(44.270)	3.735(53.424)	6.234(0.715)***
ρ			0.018(0.217)	0.024(0.360)	-0.157(1.957)	0.311(0.035)***
σ^2	9.393(0.106)***	7.26(0.059)***	1.61(14.172)	3.308(44.297)	5.353(54.414)	1.076(0.712)
ψ_0			-1.466(0.165)***	-1.408(0.192)***	-2.399(0.388)***	
ψ_1				-0.025(0.044)	-0.500(0.166)**	
ψ_2					0.584(0.125)***	

Note: * is programmed and estimated by using PROC IML, and optimized by NLPNRR routine. The standard error for each of coefficients is from 100 bootstrapping replicates, since GLLAMM in STATA did not converge after 24 hours iteration

Figure 5.1 Add Health Sampling Design



REFERENCES

1. Arellano, Manuel. (1987). Computing Robust Standard Errors for Within-Groups Estimators. Oxford Bulletin of Economics and Statistics, 49: 431-34.

2. Asparouhov, T. (2004). Weighting for Unequal Probability of Selection in Multilevel Modeling. Mplus Web Notes: No.8.

3. Asparouhov, T. (2006). General Multi-Level Modeling with Sampling Weights. Communications in Statistics - Theory and Methods, 35: 439 - 460.

4. Biemer, P., & Christ, S.L. (2007). Weighting survey data. In J. Hox, E. deLeeuw & D. Dillman (Eds.), International Handbook of Survey Methodology. Mahwah, NJ: Lawrence Erlbaum Associates.

5. Biemer, Paul P., and Sharon L. Christ. (2008). Weighting Survey Data. In International Handbook of Survey Methodology, eds. Edith, de Leeuw, Joop, J. Hox, and Don, A. Dillman. New York: Taylor & Francis Group.

6. Binder, D. A. (1983). On The Variances of Asymptotically Normal Estimators from Complex Surveys. International Statistical Review, 51(3):279 - 292.

7. Binder, D.A. (1996). Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach. Survey Methodology, 22: 17-22

8. Binder, D.A., Kovacevic, M.S., and Roberts, G. (2004). Design-Based Methods for Survey Data: Alternative Uses of Estimating Functions. Proceedings of The Section on Survey Research Methods, 3301-3312, JSM, Toronto

9. Brewer, K.R.W. and Mellor, R.W. (1973). The Effect of Sample Structure on Analytical Surveys. Australian Journal of Statistics 15: 145-152.

10. Brick, J. M., and G. Kalton. (1996). Handling Missing Data in Survey Research, Statistical Methods in Medical Research 5 (2): 2 15-38.

11. Chantala, Kim, Dan Blanchette, and C. M. Suchindran. (2006). Software to Compute Sampling Weights for Multilevel Analysis. Available online at: http://www.cpc.unc.edu/restools/data_analysis/ml_sampling_weights .

12. Christ, Sharon L., (2008), Assessment and Robust Analysis of Survey Errors, PhD dissertation.

13. Cochran WG. (1977). Sampling Techniques (3rd edn). John Wiley & Sons: New York.

14. Cook, S. R. and Gelman, A. (2006). Survey Weighting and Regression. Technical report, Department of Statistics, Columbia University.

15. Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society Series B (Statistical Methodology), 39, 1-38.

16. Diggle, P.J. and Kenward, M.G. (1994). Informative Drop-Out In Longitudinal Data Analysis, Applied Statistics, 43(1): 49-93.

17. Dippo, C. S., Fay, R. E. and Morganstein, D. H.(1984). Computing Variances from Complex Samples with Replicate Weights. Proceedings Survey Research Section of the American Statistical Association, 489-94.

18. DuMouchel, W.H. & Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association 78: 535-543

19. Eicker, F., (1967). Limit Theorems for Regressions with Unequal and Dependent Errors, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1.Berkeley: University of California Press, 59-82.

20. Eideh, A. H. and Nathan, G. (2006). Fitting Time Series Models for Longitudinal Survey Data under Informative Sampling. Journal of Statistical Planning and Inference, 136, 3052-3069.

21. Eideh, A. H. and Nathan, G. (2009). Two-Stage Informative Cluster Sampling with application in Small Area Estimation. Journal of Statistical Planning and Inference.139, 3088-3101.

22. Enders, C.K. (2001). A Primer on Maximum Likelihood Algorithms Available For Use with Missing Data. Structural Equation Modeling, 8, 128-141.

23. Fienberg, S. E. (1989). Modeling Considerations: Discussion from a Modeling Perspective. In Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M. P. (eds.), Panel Surveys, Wiley, New York. 512-539.

24. Fuller, W. A. (1975), Regression Analysis for Sample Survey, Sankhya Series C 37, 117–132.

25. Fuller, W.A. (1984). Least Squares and Related Analyses for Complex Survey Designs. Survey Methodology 10: 97-118.

26. Gelman, A. (2007). Struggles With Survey Weighting and Regression Modeling, Statistical Science, 22: 153–164

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

28. Godambe, V.P. and Thompson, M.E. (1986). Parameters of Superpopulation and Survey Population: Their Relationships and Estimation. International Statistical Review 54: 127-138.

29. Graham, J. W., Hofer, S. M., Donaldson, S. I., Mackinnon, D. P., and Schafer, J. L. (1997). Analysis with Missing Data in Prevention Research in S. West (Ed.), the

Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research. Washington D.C.: American Psychological Association. 325-366.

30. Graubard, B. I. and Korn, E. L. (2002). Inference for Superpopulation Parameters Using Sample Surveys. Statistical Science. 17: 73–96.

31. Groves, R. Fowler, F., Couper, M. Lepkowski,J, Singer, E. and Tourangeau, R. (2004). Survey Methodology, NY: Wiley.

32. Hagan, John and Holly Foster. (2003). S/He's A Rebel: Toward A Sequential Stress Theory Of Delinquency And Gendered Pathways To Disadvantage In Emerging Adulthood. Social Forces 82: 53-86

33. Hannon, Lance. (2003). Poverty, Delinquency, and Educational Attainment: Cumulative Disadvantage or Disadvantage Saturation? Sociological Inquiry 73: 575-594.

34. Harris KM, Florey F, Tabor J, Bearman PS, Jones J, Udry JR (2003): The National Longitudinal Study of Adolescent Health: Research design."

35. Harris, Kathleen Mullan, Carolyn Tucker Halpern, Pamela Entzel, Joyce Tabor, Peter S. Bearman, and J. Richard Udry. (2008). the National Longitudinal Study of Adolescent Health: Research Design. URL: http://www.cpc.unc.edu/projects/addhealth/design.

36. Haynie, Dana L. (2001). Delinquent peers revisited: Does network structure matter?, American Journal of Sociology, 106: 1013-1057.

37. Haynie, Dana L. (2003). Contexts of Risk? Explaining The Link Between Girls' Pubertal Development And Their Delinquency Involvement. Social Forces 82: 355-397.

38. Heckman JI. (1976). the Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and A Simple Estimator for Such Models. Annals of Economic and Social Measurement. 5:475–492

39. Hoem, J. (1989). The Issue of Weights in Panel Surveys of Individual Behavior. In Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M. P., editors, Panel Surveys, 512-539. Wiley.

40. Holt D, and Elliot D. (1991). Methods of Weighting for Unit Nonresponse (correction: v41, p. 599). Statistician 40:333–342.

41. Horton NJ and Kleinman KP, (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician;61: 79-90.

42. Horton, NJ & Lipsitz, SR. (2001) Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables. The American Statistician 55(3): 244-254.

43. Huber P. J., (1967), The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability , Vol. I (Berkeley: University of California Press), 221–233.

44. Iannacchione, V. G., Milne, J. G., & Folsom, R. E. (1991). Response probability weight adjustment using logistic regression. Proceedings of the American Statistical Association, 637-642

45. Ibrahim, etc (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. Journal of the American Statistical Association, Vol. 100, pp. 332-346.

46. Jeffrey M. Wooldridge, (2003). Cluster-Sample Methods in Applied Econometrics, American Economic Review, American Economic Association, vol. 93(2), 133-138, May.

47. Kalsbeek, William D., Juan Yang, and Robert P. Agans, (2002). Predictors of Nonresponse in a Longitudinal Survey of Adolescents. Proceedings of the 2002 American Statistical Association, Survey Research Methods Section

48. Kalton, G. (1983). Compensating for Missing Survey Data. Survey Research Center, University of Michigan, Ann Arbor, Michigan.

49. Kalton, G. (1989). Modeling Considerations: Discussion from a Survey Sampling Perspective. In Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M., editors, Panel Surveys, 575-585. Wiley.

50. Kalton, G. and Kasprzak, D. (1986) the Treatment Of Missing Survey Data. Survey Methodology, 12, 1-16.

51. Kish, L. (1965). Survey Sampling. John Wiley and Sons, New York.

52. Kish, L. and Frankel, M. R. (1974). Inference from Complex Samples, Journal of the Royal Statistical Society, Series B 36: 1–37.

53. Korn, E. and Graubard, B. (1999). Analysis of Health Surveys. John Wiley & Sons, Inc.

54. Korn, E. L. and Graubard, B. I. (2003). Estimating Variance Components by Using Survey Data. Journal of the Royal Statistical Society, Series B, 65(1):175 - 190.

55. Korn, E.L., and Graubard, B.I. (1995). Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. The American Statistician, 49: 291-295.

56. Krieger, A. M. and Pfeffermann, D. (1992). Maximum likelihood from complex sample surveys. Survey Methodology, 18: 225-239.

57. Krieger, A. M. and Pfeffermann, D. (1997). Testing of Distribution Functions from Complex Sample Surveys. Journal of Official Statistics, 13: 123-142.

58. Laird, N. M. (1988) Missing data in longitudinal studies. Statistics in Medicine, 7: 305-315.

59. Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data, Biometrics 38: 963–974.

60. Lehtonen, R. and E. Pahkinen (2004). Practical Methods for Design and Analysis of Complex Surveys, John Wiley & Sons

61. Liang, K. Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. Biometrika 73: 13–22.

62. Lin, D. Y. and L. J. Wei. (1989). the Robust Inference for the Cox Proportional Hazards Model. Journal of the American Statistical Association, 84: 1074-1078.

63. Little R.J.A. (1993). Pattern-Mixture Models for Multivariate Incomplete Data. Journal of the American Statistical Association. 88:125–34

64. Little, R. (1993). Post-stratification: A Modeler's Perspective. Journal of the American Statistical Association, 88:1001-1012.

65. Little, R. (2004). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling. Journal of the American Statistical Association, 99:546-556.

66. Little, R. J. A. (1992) Regression With Missing X's: A Review. Journal of the American Statistical Association, 87: 1227-1237.

67. Little, R.J.A. & Rubin, D.B. (1987) Statistical Analysis with Missing Data. New York, Wiley.

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

69. Lohr, S. and Liu, J. (1994). A Comparison of Weighted and Unweighted Analyses in the ncvs. Journal of Quantitative Criminology, 10:343-360

70. Lohr, S. L. (1999). Sampling: Design and Analysis. Duxbury Press.

71. Mislevy, R. J. and Sheehan, K. M. (1989). The Role of Collateral Information about Examinees in Item Parameter Estimation. Psychometrika, 54(4):661-679.

72. Molenberghs G, and etc. (2007), Analysis of Incomplete Data. In: Pharmaceutical Statistics with SAS, Alex Dmitrienko, Christy Chuang-Stein, Ralph D'Agostino (edts). SAS Publishing.

73. Nathan, G. & Holt, D. (1980). The Effect of Survey Design on Regression Analysis. Journal of the Royal Statistical Society, Ser. B 42: 377-386.

74. Nordberg, L. (1989). Generalized Linear Modeling of Sample Survey Data. Journal of Official Statistics 5: 223-239.

75. Patterson, B., Dayton, M., and Graubard, B. (2002). Latent Class Analysis of Complex Sample Data: Application to Dietary Data. Journal of the American Statistical Association, 97(459): 1-21.

76. Pfeffermann D. (1996). The Use of Weights in Survey Analysis. Statistical Methods in Medical Research; 5: 239-261

77. Pfeffermann, D. & Holmes, D.J. (1985). Robustness Considerations in the Choice of Method of Inference for Regression Analysis of Survey Data. Journal of the Royal Statistical Society, Ser. A 148: 268-278.

78. Pfeffermann, D. (1993). The Role of Sampling Weights When Modeling Survey Data. International Statistical Review, 61(2):317-337.

79. Pfeffermann, D. and Holmes, D.J. (1985). Robustness Considerations in the Choice of Method of Inference for Regression Analysis of Survey Data. Journal of the Royal Statistical Society, Ser. A 148,268-278

80. Pfeffermann, D., Krieger, A.M., and Rinott, Y. (1998). Parametric Distributions of Complex Survey Data under Informative Probability Sampling. Statistica Sinica, 8, 1087-1114

81. Pfeffermann, D., Moura, F. A. D. S., and Silva, P. L. D. N. (2006). Multilevel Modeling under Informative Sampling. Biometrika, 93:943-959.

82. Pfeffermann, D., Skinner, C., Goldstein., Holmes, D., and Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models (with discussion). Journal of the Royal Statistical Society, Series B, 60, 23-56.

83. Rabe-Hesketh, S. and Skrondal, A. (2006), Multilevel Modeling of Complex Survey Data, Journal of the Royal Statistical Society, Series A, 169, 805–827.

84. Raghunathan T.E. (2004). What Do We Do With Missing Data? Some Options for Analysis of Incomplete Data. Annual Review of Public Health, 25, 99-117.

85. Rao, J. N. K J. G. Kovar & H. J. Mantel (1990). On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information. Biometrika , 77: 365-375.

86. Rao, J. N. K., and Shao, J. (1999). Modified Balanced Repeated Replication for Complex Survey Data. Biometrika, 86: 403-415.

87. Rao, Poduri S. R. S. (2000), Sampling Methodologies with Applications. Chapman and Hall/CRC Press, London and New York.

88. Rizzo L., Kalton G., Brick J.M. (1996) A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse, Survey Methodology, 22, 43-53.

89. Royall, R. M. (1986). Model Robust Confidence Intervals Using Maximum Likelihood Estimators. International Statistical Review, 54: 221-226.

90. Rubin DB, Stern H, Vehovar V. (1995). Handling “Don’t Know” Survey Responses: The Case of Slovenian Plebiscite. Journal of the American Statistical Association. 90:822–28.

91. Rubin DB. (1974). Characterizing the Estimation of Parameters in Incomplete Data Problems. Journal of the American Statistical Association. 69:467–474.

92. Rubin DB. (1977). Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys. Journal of the American Statistical Association.72:538–543.

93. Rubin DB. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley

94. Rubin, D.B. (1976) Inference and Missing Data. Biometrika, 63, 581-592

95. Rubin, D.B. (1985). The Use of Propensity Scores in Applied Bayesian Inference. Bayesian Statistics 2, eds.J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith, Elsevier Science Publishers B.V. pp.463-472.

96. Rubin, D.B. (1996) Multiple Imputation after 18+ Years (with discussion). Journal of the American Statistical Association, 91, 473-489.

97. Rubin, D.B. (2003), Nested Multiple Imputation of NMES via Partially Incompatible MCMC, Statistica Neerlandica, 57, 3-18.

98. Rubin, D.B., and Schenker, N. (1991), Multiple Imputation in Health-Care Data Bases: An Overview and Some Applications, Statistics in Medicine, 10, 585-598

99. Rust, K. (1985). Variance Estimation for Complex Estimators in Sample Surveys. Journal of Official Statistics. 1(4): 381-397.

100. Sarndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag.

101. Searle, S. R., G. Casella, and C. E. McCulloch. (1992). Variance Components. New York: Wiley.

102. Siniff, D. B., and R. O. Skoog. (1964). Aerial Censusing of Caribou Using Stratified Random Sampling. Journal of Wildlife Management. 28:391-401.

103. Skinner, C J, Holt, D and Smith, T M F, eds (1989) Analysis of Complex Surveys. Chiehester: Wiley.

104. Skinner, C. J. (1989). Domain Means, Regression and Multivariate Analysis. In Skinner, C. J.,Holt, D., and Smith, T. M. F., editors, Analysis of Complex Surveys, pages 59-87. Wiley.

105. Skinner, C. J. (1994). Sample Models and Weights. In Proceedings of the Section on Survey Research Methods, 133-142. American Statistical Association, Alexandria, VA.

106. Steel, R. G. D., and J. H. Torrie. (1980). Principles and Procedures of Statistics. Second ed. McGraw-Hill, New York, N.Y.

107. Thomas, D. R. and Cyr, A. (2002). Applying Item Response Theory Methods to Complex Survey Data. In Proceedings of the Survey Methods Section, 17-25. Statistical Society of Canada.

108. Tourangeau, R. and H. Shin (1999). National Longitudinal Study of Adolescent health: Grand Sample Weight. National Opinion Research Center and Carolina Population Center. website: http://www.cpc.unc.edu/projects/addhlth/.

109. Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. New York: Springer

110. Vermunt, J. and Magidson, J. (2007). Latent class analysis with sampling weights, a maximum likelihood approach. Sociological Methods & Research, 36(1):87-111.

111. White, H., (1980), A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, Econometrica 48, 817-838.

112. Wolter, Kirk M. (1985). Introduction to Variance Estimation. Springer-Verlag: New York.

113. Woodruff, R. A. (1971). Simple Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association, 66, 411-414.