CARNEGIE MELLON UNIVERSITY

TO WEIGHT OR NOT TO WEIGHT? INCORPORATING SAMPLING DESIGNS INTO MODEL-BASED ANALYSES

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY
In
STATISTICS

by

MARIANNE (MARNIE) BERTOLET

Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

July, 2008

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

Carnegie Mellon

COLLEGE OF HUMANITIES & SOCIAL SCIENCES

DISSERTATION

Submitted in Partial Fulfillment of the Requirements

For the Degree of **DOCTOR OF PHILOSOPHY**

Title:

" TO WEIGHT OR NOT TO WEIGHT: INCORPORATING SAMPLING DESIGNS INTO MODEL-BASED ANALYSES "

Presented by: **MARIANNE BERTOLET**

Accepted by the DEPARTMENT OF STATISTICS

Readers:

(Director of Dissertation)	22 July 2008 Date
Howard Seltman	7/21/08 Date
Joel Greenhouse	7/22/08 Date
	7/22/08 Date
David Banks	2008/7/22 Date

Approved by the Committee on Graduate Degrees

John Lehoczky (Dean)	8/12/08 (Date)

Abstract

Large-scale statistical surveys seldom use simple random sampling. Two fundamental approaches (design- and model-based) exist to incorporate complexities such as stratification, clustering and/or unequal probabilities of selection into the survey analysis. The debate about design- vs. model-based analysis has a long history and currently centers around the role of design-based sampling weights in model-based analyses. This thesis begins by investigating three different published proposals on how to insert the sampling weights into linear mixed-effects (LME) models. This component, which focuses on *how* the sampling weights are inserted into LME models, derives the three methods from a common starting place and emphasizes the unique decisions that distinguish the different approaches. The second component of this thesis compares the methods in a simulation study that varies the types of informative sampling and model misspecification. The goal of this component is to characterize *when* it is appropriate to include sampling weights into a model-based analysis, as well as *which* kinds of sampling and modeling errors weights can correct. Finally, the lessons from the first two components are extended to the Grade of Membership (GoM) model, a hierarchical Bayesian mixed-membership model whose variance components do not map well to the dependencies induced by complex sampling designs. The GoM model is modified to include a polytomous logistic mixed-effects regression prior to reflect the sampling design. A new type of weighting, called *weighting based on the estimated parameter* is developed and explored through a simulation study.

Acknowledgements

I would like to thank my husband, Mark Bertolet, for supporting my decision to leave a comfortable industry job to pursue my doctoral degree. He has supported and encouraged me through all the high and low points. Along with my daughters Elise and Claire, Mark reminds me what is important and how to remain true to myself.

I would like to thank my advisor, Brian Junker. Brian has taught me many things, most importantly how to approach new problems and how to write coherently about them. These are two skills that I will keep with me and use in all my future endeavors. I thank him for his patience, his insight, and his constant reminding about the differences between effect and affect, strata and stratum. I also thank Joel Greenhouse and Howard Seltman for starting my curiousity regarding survey sampling and for Steve Fienberg's introduction to the distinction between design- and model-based analyses.

I would like to thank the Statistics department faculty and staff at large, for supporting me as my family expanded and my priorities changed. The flexibility, understanding and guidance provided by many faculty and staff were invaluable to my continuing through the program and remaining sane at the same time. I thank all the friends who have supported me through the years. I especially thank my officemate of five years, Hoa Nguyen, whose wonderful sense of humor helped lighten difficult times. I am also indebted to Rhiannon Weaver, for holding my hand through many of the professional and personal challenges I have faced. Everyone needs friends like Hoa and Rhiannon.

Contents

List of Tables

List of Figures

Chapter 1

Introduction to Survey Sampling

1.1 Introduction

Large-scale statistical surveys seldom use simple random sampling. Two fundamental approaches (design- and model-based) exist to incorporate complexities such as stratification, clustering and/or unequal probabilities of selection into the survey analysis. Design-based analysis first originated around the early 1930's with Neyman (1934) and Fisher (1935). Model-based analysis began in the mid-1950's to mid-1960's with work by Godambe (1955), Godambe (1966) and Royall (1968). In the 1970's and 1980's, there were debates on the use and validity of the two types of analysis with Royall (1976), Hoem (1989), and Fienberg (1980, 1989) on the model-based side and Kalton (1968, 1989) and Hansen and Tepping (1983) on the design-based side, to name a few. Research from the 1990's continuing to today investigates when and how to combine the two approaches, for example see Little (1991), Pfeffermann (1993), Pfeffermann et al. (1998), Graubard and Korn (1995) and Korn and Graubard (2003). The survey world has been polarized regarding design and model-based analysis, and current research focuses on understanding how and when to combine them. The goal of this dissertation is to investigate the use of design-based sampling weights in model-based analysis. This dissertation contains three major components.

The first component investigates the theory of linear mixed-effects models with complex survey sampling data. Research to date considers the estimation of linear mixed-effects model parameters (both fixed components and variance components) using both model- and design-based analysis; see Pfeffermann et al. (1998), Korn and Graubard (2003), Asparouhov (2006), and Rabe-Hesketh and Skrondal (2006). While these methodologies merge design- and model-based methodologies in one analysis, they lack guiding principles (Little, 2004).

This component focuses on deriving three published and competing approaches for adding design-based weights to the merged design- and model-based analyses, with all derivations starting at a common spot and emphasizing where the approaches make unique decisions. These derivations emphasize the ad-hoc nature of the approaches. This component focuses on *how* to incorporate weights into mixed-effects models.

The second component investigates the properties of the different approaches through a comprehensive simulation study. The simulations also compare different types of weights and their affects on linear mixed-effects models with various types of model misspecification and informative sampling. This component focuses on *when* to incorporate weights into mixed-effects models and to a lesser extent *which weights* to incorporate.

The third component extends the lessons from weighting the linear mixed-effects models to a Bayesian formulation of the Grade of Membership Model (GoM) (Erosheva, 2002). The results are demonstrated through simulations, and future work includes applying the expanded GoM model to data from the National Long Term Care Survey, NLTCS (1988).

The remainder of this chapter provides necessary background for the thesis. This includes information on survey sampling, linear models and extensions to more complex models.

1.2 Survey Sampling

The complexities in survey sampling data revolve around the sampling design. The sampling design often induces dependencies and creates samples that are not representative of the population. Of specific concern is when the sampling is informative, or related to the outcome of interest. Longford (2004) defines an informative sampling design with respect to a set of covariates X as a design where the outcomes and the inclusion probabilities are correlated even after conditioning on X, i.e. $\text{Cov}(Y, \pi|X) \neq 0$, where Y represents the outcomes and π represents the probabilities that the elements are included in the sample. Note the similarity between the definitions of informative sampling and informative missing data. These similarities are noted in Rubin (1976) and expanded upon in Sugden and Smith (1984) and Smith and Sugden (1988). As such, the missing data terminology from Little and Rubin (2002) can be adapted to sampling designs. Define the levels of informative sampling as follows:

Sampling Completely At Random (SCAR): *Unconditional* on X, Y and π are *independent*, $\text{Cov}(Y, \pi) = 0$,

Sampling at Random, (SAR): *Conditional* on X, Y and π are *independent*, $\text{Cov}(Y, \pi|X) = 0$ for all observed Y and X and

Not Sampling At Random (NSAR), or Informative Sampling: *Conditional* on X, Y and π are still *dependent*, $\text{Cov}(Y, \pi|X) \neq 0$.

With these definitions of different types of informative sampling, we next look at various sampling designs. The sampling design refers to the way in which the sample is chosen. Sampling designs may have many levels comprised of strata and/or clusters.

In *stratified sampling*, the population is partitioned into H subpopulations (or strata) that do no overlap and that encompass the entire population. Independent probability samples are drawn from each stratum. Stratified samples ensure adequate representation from each of the H strata and protect against bad samples.

In *cluster sampling*, the population is partitioned into K clusters that do not overlap and that encompass the entire population. A sample of clusters is taken, and only elements in the sampled clusters are candidates for inclusion in the sample. Cluster sampling reduces costs as the sample is only administered to individuals in sampled clusters, who are usually geographically close or otherwise easy to sample together. Cluster sampling, however, induces dependencies between the elements in the cluster.

Multilevel sampling uses multiple levels of stratification and clustering. An example of a multilevel sampling design would be to stratify, then cluster then stratify. For example, in a national survey we may want to stratify on state (to ensure representation from each state), cluster on county (to reduce travel within each state) and then stratify on income level (to ensure representation from each income level).

When analyzing survey sample data, the sampling design needs to be considered. Historically, the sampling design is incorporated into the analysis by using design-based methods. Recently, model-based methods and hybrid model- and design- based methods are becoming popular. The different methods are discussed in the next section. The key concepts for these methods are described in this section and are summarized into the following major categories:

1. The population or estimand(s) about which we wish to make inference;
2. The source(s) of variability in the data we observe;
3. The role(s) of sampling weights.

A summary of these key differentiating points and their implications is in Table 1.1. These differences are explored further in the next sections.

1.2.1 Design-Based Analysis

The key differentiating points for design-based analysis are: 1) the population of interest is the specific finite population that was sampled, 2) the variability is induced by the sampling

	Design-Based Analysis	Model-Based Analysis	Hybrid Analysis
Data	Fixed	Random	Random
Origin of variability	Difference across samples	Model error term	Both differences in samples and model error term
Probability distribution	Randomization distribution	Model distribution	Joint randomization and model distribution
Target of inference	Finite population	Underlying stochastic model or finite population	Superpopulation inference
If a census is observed	Parameters known exactly	Model parameter not known exactly Finite population parameter known exactly	Superpopulation parameters not known exactly
Sampling design	Define underlying distribution	Identify dependencies and additional variation	Partially identify dependencies and additional variation
Sampling weights	Needed	Controversial	Needed for randomization distribution

Table 1.1: Summary of Design- vs. Model-Based Methodologies

design and 3) the sampling weights are crucial for the analysis.

More formally, let $\mathcal{U}$ represent the finite population of N elements, and $\beta_{\mathcal{U}}$ be the target parameter to be estimated. $\beta_{\mathcal{U}}$ is a summary statistic, sometimes called a descriptive population quantity, for the finite population. This target parameter is a function of the census data and must be estimated from a sample of n elements. For a given sampling design, $p_{\mathcal{U}}(.)$, suppose there are t_n possible samples of size n from the population. Let $p_{\mathcal{U}}(s)$ be the probability of selecting the s^{th} sample and define $\hat{\beta}_{p_{\mathcal{U}}}^{(s)}$ to be the estimate of $\beta_{\mathcal{U}}$ from the s^{th} sample. Then $\{(\hat{\beta}_{p_{\mathcal{U}}}^{(s)}, p_{\mathcal{U}}(s))\}_{s=1}^{t_n}$ defines the randomization distribution (Lohr, 1999). If all the elements in the finite population are included in the sample, then there is no variability in the randomization distribution and the finite population quantity is known exactly. The randomization distribution is also referred to as the sampling distribution.

As defined above, the estimates of the parameter from all the possible samples from the population are needed to know the randomization distribution. Sampling weights provide information on the non-sampled data. Let I_{hi} be a random indicator variable taking on

the value 1 if the hi^{th} element is included in the sample, and 0 otherwise. Note that the hi subscripts will vary according to the sampling design. Note that $E(I_{hi}) = \pi_{hi} = P(I_{hi} = 1)$, where π_{hi} is the probability, under $p_{\mathcal{U}}$, that the hi^{th} element is included in the sample. The sampling weight is defined as $w_{hi} = \frac{1}{\pi_{hi}}$ (Korn and Graubard, 1999). An interpretation of w_{hi} is that it corresponds to the number of people that element hi represents in the population. These weights are important for two reasons; 1) in all but the simplest of sampling designs, design-based analyses without the weights have very large bias and 2) the design-based analysis gains information about the non-realized samples from the sampling weights and the sampling structure.

As an example of a design-based analysis, assume a stratified sampling design with H strata, and within each stratum there is simple random sampling (the sampling design is SCAR if the sample proportions for each stratum match the population proportions, and SAR if there is disproportionate sampling). The i^{th} element in the h^{th} stratum has an associated fixed quantity y_{hi} and random indicator variable I_{hi}. Let $\mathcal{U}_h$ represent the population in the h^{th} strata. Assume the population has N population elements, that stratum h has N_h population elements of which n_h are sampled. Suppose that target of interest is the mean of the finite population, $\bar{y} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h} y_{hi}$. A common estimator for the mean is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), $\hat{\bar{y}}_{HT}^{(D)}$. The superscript (D) indicates this is a design-based estimate and the subscript HT indicates the Horvitz-Thompson estimator. This estimator is a weighted average of the sample elements,

$$\hat{\bar{y}}_{HT}^{(D)} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h} w_{hi} I_{hi} y_{hi}.$$

Though the sum in the Horvitz-Thompson estimator appears to be over the entire population, the random inclusion indicator variables only allow elements in the sample to add to the sum term. This estimator is unbiased, because $E(I_{hi}) = \frac{1}{w_{hi}}$ (y_{hi} is fixed). The covariance between the random indicator variables is needed to compute the variance. Lohr

(1999) derives

$$\mathrm{Var}(\hat{\bar{y}}_{HT}^{(D)}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\mathrm{S}_h^2}{n_h}.$$

In the above equation, S_h^2 represents the sample variance of the sampled elements in the h^{th} stratum (Sarndal, Swensson and Wretman, 1992). The $1-\frac{n_h}{N_h}$ term is often referred to as the finite population correction. Note that if a census is taken and $n_h = N_h$ then the variance becomes zero.

In large complex surveys, the sampling weights provided with the data are rarely inverse probability weights. The sampling weights are adjusted for many reasons, as described in Chapter 4 of Korn and Graubard (1999). Some common adjustments are for non-response, misspecification of the sampling frame, poststratification and missing data. In the remainder of this thesis, it is assumed that the sampling weights are inverse probability weights. Future work is needed to determine the effect of the adjustments on analysis.

1.2.2 Model-Based Analysis

The key differentiating points for model-based analysis are: 1) the population of interest is either the specific finite population that was sampled or a generating model parameter (or superpopulation parameter), 2) the variability is induced by a stochastic model assumed to generate variables associated with each element of the finite population and by the sampling design and 3) the use sampling weights is very controversial.

In model-based analysis, assume that for element hi there is an associated value Y_{hi} that was randomly generated by a model ξ. Consider this as a two-stage process, where the first stage generated the finite population of size N from the ξ model, and the second stage sampled the finite population. In this scenario, the estimation can be to a finite population parameter, or a superpopulation parameter. A superpopulation estimand refers to a parameter (or function of parameters) from the generating stochastic model.

Both frequentist and Bayesian model-based approaches are introduced in this chapter. Later chapters in this thesis use frequentist linear mixed-effects models and Bayesian grade of membership models. The frequentist approach is based on methods such as maximum likelihood estimation where the likelihood is based on both the sampling design and the generating model. One Bayesian approach takes the census likelihood, and integrates out the non-sampled values and makes inference based on posterior distributions. This approach is used in the example below to show how a Bayesian model-based estimator can match the design-based estimator. In Chapter 4 of this thesis, the Bayesian approach is derived by using an estimate of the census likelihood to form the posterior and does not use the missing data framework.

Frequentist Model-Based Analysis

As an example of frequentist model-based analysis, consider the superpopulation average, that is, the expected value of the average of any similarly constructed population created by the stochastic mechanism. Assume the same sampling design as the design-based example. Let the data be generated as $Y_{hi} = \mu_k + \epsilon_{hi}$ where $\epsilon_{hi} \sim N(0, \sigma_h^2)$. The parameter of interest is $\bar{Y} = \frac{1}{N}\sum_{h=1}^{H} N_h \mu_h$. To estimate $\bar{Y}$, each μ_h needs to be estimated. We know that the MLE for the mean of each stratum is the sample mean for the stratum. Inserting the MLE for μ, we get

$$\begin{aligned} \hat{\bar{Y}}^{(FM)} &= \frac{1}{N}\sum_{h=1}^{H} N_h \sum_{i \in \mathcal{U}_h} \frac{1}{n_h} I_{hi} Y_{hi} \\ &= \frac{1}{N}\sum_{h=1}^{K} \sum_{i \in \mathcal{S}_h} w_{hi} Y_{hi} \end{aligned}$$

where the superscript (FM) represents the frequentist model-based estimator. The weights are included in this model-based analysis, not because they were inserted for the randomization distribution, but because they appeared due to numerical computations ($w_{hi} = N_h/n_h$

because of the simple random sampling). This estimator matches the design-based estimator, though it was derived under different assumptions. When taking the expected value of this estimator, both the distribution of the Y variables and the distribution of the I variables are considered. As a result

$$E(\hat{\bar{Y}}^{(FM)}) = E_\xi E_p(\frac{1}{N}\sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h} w_{hi}I_{hi}Y_{hi}) = E_\xi(\frac{1}{N}\sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h} Y_{hi}) = \frac{1}{N}\sum_{h=1}^{H} N_h\mu_h. \quad (1.1)$$

The variance is computed by conditioning on the finite population, Y_C,

$$\begin{aligned}\mathrm{Var}(\hat{\bar{Y}}^{(FM)}) &= \mathrm{Var}_\xi(E_p(\hat{\bar{Y}}^{(FM)}|Y_C)) + E_\xi(\mathrm{Var}_p(\hat{\bar{Y}}^{(FM)}|Y_C)) \\ &= \frac{1}{N^2}\sum_{h=1}^{H} N_h\sigma_h^2 + \frac{1}{N^2}\sum_{h=1}^{H} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}. \quad (1.2)\end{aligned}$$

The first term of the variance is the variance if a census were taken, and the second term of the variance is the added variance because a census was not taken. When a census is taken, the second term becomes zero.

Bayesian Model-Based Analysis

For the Bayesian model-based approaches prior distributions are placed on the parameters of the model. All inference then is with respect to the posterior distributions of the parameters. An early example is found in Sugden (1979) and Sugden (1985) discusses ignorable design in context of Bayesian analysis. Gelman (2007) summarizes current status of Bayesian analysis and survey weighting. One Bayesian model-based approach treats the non-sampled elements as missing values that are integrated out of the likelihood, (Gelman et al., 2004). Links between Bayesian analysis and design-based analysis have been studied by Holt and Smith (1979) and Little (1991, 1993), to name a few.

As an example of Bayesian model-based analysis, consider the finite population average. This example differs from the others for two reasons: 1) the use of Bayesian methodologies

instead of frequentist methodologies and 2) the use of a finite population estimand instead of a superpopulation estimand. The sampling design and generating stochastic mechanism are the same as the previous frequentist examples. Let Y_C, Y_S and $Y_{C\setminus S}$ represent the census, sampled and non-sampled elements respectively. Then

$$\bar{Y} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h} Y_{hi} = \frac{n}{N}\bar{Y}_S + \frac{N-n}{N}\bar{Y}_{C\setminus S}.$$

Assume that $\mu_h \sim N(\mu_{0h}, \sigma^2_{0h})$ where μ_{0h} and σ^2_{0h} are known and the μ_h's are independent. The census likelihood equation is

$$p(Y_C, I|\mu, \sigma^2) = \prod_{h=1}^{H}\prod_{i\in\mathcal{U}_h} \phi(Y_{hi}|\mu_h, \sigma^2_h)p(I_{hi}),$$

where $\phi(Y_{hi}|\mu_h, \sigma^2_h)$ refers to the value of the normal density, with mean μ_h and variance σ^2_h, at Y_{hi}. This assumes that the sampling mechanism is independent of the data, or that the sampling is SCAR. In addition, the SRS in each stratum is independent, so $p(\{I_{hi} = 1\}) = \frac{n_h}{N_h}$. The sample likelihood can be obtained by integrating non-sampled elements,

$$\begin{aligned} p(Y_S, I|\mu, \sigma^2) &= \int \prod_{h=1}^{H}\left[\left(\prod_{i\in\mathcal{S}_h}\frac{n_h}{N_h}\phi(Y_{hi}|\mu_h, \sigma^2_h)\right)\left(\prod_{j\notin\mathcal{S}_h}\frac{N_h-n_h}{N_h}\phi(Y_{hj}|\mu_h, \sigma^2_h)\right)\right] dY_{C\setminus S} \\ &\propto \prod_{h=1}^{H}\prod_{i\in\mathcal{U}_h} I_{hi}\phi(Y_{hi}|\mu_h, \sigma^2_h). \end{aligned}$$

Next, the posterior mean is calculated as

$$\begin{aligned} p(\mu|Y_S, I) &\propto p(\mu)p(Y_S, I|\mu) \\ &\propto \left(\prod_{h=1}^{H} \phi(\mu_h|\mu_{0h}, \sigma_{0h}^2)\right)\left(\prod_{h=1}^{H}\prod_{i\in\mathcal{U}_h} I_{hi}\phi(Y_{hi}|\mu_h, \sigma_h^2)\right) \\ &= \prod_{h=1}^{H}\left(\phi(\mu_h|\mu_{0h}\sigma_{0h}^2)\prod_{i\in\mathcal{U}_h} I_{hi}\phi(Y_{hi}|\mu_h, \sigma_h^2)\right) \\ &= \prod_{h=1}^{H} p(\mu_h|Y_S, I). \end{aligned}$$

Let $\bar{Y}_{S_h}$ be the sample average of the Y values in the h^{th} stratum. The posterior μ_h distributions are independent and normally distributed, and using the standard Bayesian conjugacy, they have the following mean and variance

$$\begin{aligned} E(\mu_h|Y_S, I) &= \bar{Y}_{S_h}\left(\frac{n_h\sigma_{0h}^2}{n_h\sigma_{0h}^2+\sigma_h^2}\right) + \mu_{0h}\left(\frac{\sigma_h^2}{n_h\sigma_{0h}^2+\sigma_h^2}\right) \\ \text{Var}(\mu_h|Y_S, I) &= \frac{\sigma_h^2\sigma_{0h}^2}{n_h\sigma_{0h}^2+\sigma_h^2}. \end{aligned}$$

Note that the posterior distribution of $\frac{1}{N}\Sigma_{h=1}^{H} N_h\mu_h$ would be used for Bayesian inference for the superpopulation parameter. To get the finite population estimate, the posterior predictive distribution is needed. We know that $p(\bar{Y}_{C\setminus S}|\mu, I)$ is normally distributed with mean $\frac{1}{N-n}\Sigma_{h=1}^{H}(N_h - n_h)\mu_h$ and variance $\frac{1}{(N-n)^2}\Sigma_{h=1}^{H}(N_h - n_h)\sigma_h^2$. We showed above that $p(\mu|Y_S, I)$ is normally distributed, so $p(\bar{Y}_{C\setminus S}|Y_S, I)$ is approximately normally distributed under the SCAR design,

$$p(\bar{Y}_{C_h\setminus S_h}|Y_{S_h}, I) \sim \text{Normal} \quad \text{Mean} \approx \bar{Y}_{S_h} \quad \text{Variance} \approx \frac{N_h}{(N_h - n_h)}\frac{\sigma_h^2}{n_h}$$

See Gelman et al. (2004) for details. Recall the goal is to estimate $\bar{Y}_C = \frac{n}{N}\bar{Y}_S + \frac{N-n}{N}\bar{Y}_{C\setminus S} = \frac{1}{N}\sum_{h=1}^{H} N_h\left(\frac{n_h}{N_h}\bar{Y}_{S_h} + \frac{N_h-n_h}{N_h}\bar{Y}_{C/S_h}\right)$. Because $\bar{Y}_S$ is considered a constant to the posterior

distribution, we can easily find the posterior distribution of $\bar{Y}_C$.

$$\bar{Y}_C \sim \text{Normal} \quad \text{Mean} \approx \bar{Y}_S \quad \text{Variance} \approx \frac{1}{N^2}\sum_{k=1}^{K} N_k^2(1-\frac{n_k}{N_k})\frac{\sigma_k^2}{n_k}$$

Note that the mean and variance of this posterior distribution match the mean and variance of the design based estimator.

1.2.3 Hybrid Analysis

Numerous hybrid approaches that combine both model- and design-based analyses have been created, summarized in Pfeffermann (1993). An early example of the hybrid analysis is in Hartley and Sielken (1975) for estimation of finite population quantities using a superpopulation model. Smith (1988) pondered the same question as I do in my title, "To Weight or Not To Weight", regarding hybrid analyses. An example of the hybrid analysis for estimation of superpopulation model parameters is in Molina et al. (2001). A recent summary of the hybrid approach is in Binder and Roberts (2006).

For this method, assume that a census is available and construct the desired estimate of the model-based parameter with the census data. Next estimate the census model-based estimator using design-based techniques. Consider again the superpopulation average, that is, the expected value of the average of any similarly constructed population created by the stochastic mechanism. Let the sampling design and generating stochastic mechanism be the same as it was in the frequentist and Bayesian model-based examples. The estimand is $\bar{Y} = \frac{1}{N}\sum_{h=1}^{H} N_h\mu_h$. We know that in the frequentist analysis, the MLE for μ_h is the mean of the data, in this case, the mean of the census values in stratum h. Substituting this in, we get $\hat{\bar{Y}}_C^{(HA)} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h} Y_{hi}$, where the superscript (HA) refers to the hybrid approach. Next, estimate this using design-based methods. This is exactly what was done in the design-based section above. The hybrid-approach estimator based on the sample

data becomes,

$$\hat{\bar{Y}}^{(HA)} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in S_h} w_{hi}Y_{hi}.$$

The weights were explicitly added because of the randomization distribution. The expected value and variance of this estimate is evaluated with respect to both the p and ξ distributions. This HA estimator is the is same as the FM estimator. Because the evaluation of the estimators is with respect to the same distributions, the expected value and variance are the same as in Equations 1.1 and 1.2.

$$\begin{aligned} E(\hat{\bar{Y}}^{(HA)}) &= \frac{1}{N}\sum_{h=1}^{H} N_h\mu_h \\ \mathrm{Var}(\hat{\bar{Y}}^{(HA)}) &= \frac{1}{N^2}\sum_{h=1}^{H} N_h\sigma_h^2 + \frac{1}{N^2}N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h} \end{aligned}$$

In this example the hybrid-approach and frequentist model-based estimators are the same. This only occurs in simple examples.

1.2.4 Evaluation Criteria

Model-based survey sampling consists of two stages, first creating the finite population from the generating stochastic mechanism, and secondly obtaining a sample from the finite population. As seen in previous sections, the criteria to evaluate the estimators becomes complicated in model-based analysis because of the need to take into account both the randomization distribution and the generating model distribution.

First consider the definition of these common criteria in the standard statistical theory. The bias of an estimator $\hat{\theta}$ for θ is $E(\hat{\theta})-\theta$, and $\hat{\theta}$ is unbiased if the bias is zero. The variance of $\hat{\theta}$ is $E((\hat{\theta}-E(\hat{\theta}))^2)$ and the mean squared error (MSE) is $E((\hat{\theta}-\theta)^2)$. The definitions of these quantities in survey sampling are very similar, as seen below. Now consider some asymptotic properties. Specifically, suppose that θ is estimated by $\hat{\theta}_n$, where $\hat{\theta}_n$ is

a function of n independent and identically distributed random variables, say $\tau_1, \tau_2, ..., \tau_n$. Then $\hat{\theta}_n$ is asymptotically unbiased if $\lim_{n\to\infty} \hat{\theta}_n = \theta$, and it is consistent if, for any fixed ϵ, $\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$. In survey sampling the finite population size is fixed at N, and the sampling size $n < N$. Adjustments to these criteria are needed to apply them to survey data.

Consider the following notation used in Hansen and Tepping (1983) and Sarndal et al. (1992). Let $y_1, y_2, y_3, ...$ be a sequence of variables of interest from an associated sequence of elements in the population, labeled $k = 1, 2, 3, ...$ A corresponding sequence of populations, $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3, ...,$ can be defined, where $\mathcal{U}_v$ consists of the first N_v elements from the infinite sequence of elements. In this construction $\mathcal{U}_1 \subset \mathcal{U}_2 \subset \mathcal{U}_3, ...$ and $N_1 < N_2 < N_3 < ....$ Let $\theta_{\mathcal{U}_v}$ be the value of a parameter in population $\mathcal{U}_v$. Each population $\mathcal{U}_v$ has a sampling design $p_{\mathcal{U}_v}()$ that assigns probability $p_{\mathcal{U}_v}(s_v)$ to sample s_v.

Assume the sample sizes are fixed at n_v, where $n_1 < n_2 < n_3 < ....$ Let $\hat{\theta}_{p_{\mathcal{U}_v}}$ be an estimator of $\theta_{\mathcal{U}_v}$ based on the sampling design $p_{\mathcal{U}_v}$. Hansen and Tepping (1983) and Sarndal et al. (1992) define the following:

- p-bias and p-variance – An estimator $\hat{\theta}_{p_{\mathcal{U}_v}}$ is said to be *p-unbiased*, or design unbiased, for a parameter $\theta_{\mathcal{U}_v}$ in population $\mathcal{U}_v$ if $E_{p_{\mathcal{U}_v}}(\hat{\theta}_{p_{\mathcal{U}_v}}) = \sum_{s_v \in \mathcal{U}_v} \hat{\theta}_{p_{\mathcal{U}_v}}(s_v) p_{\mathcal{U}_v}(s_v) = \theta_{\mathcal{U}_v}$, where $s_v \in \mathcal{U}_v$ represents all possible samples in the finite population $\mathcal{U}_v$ and $\hat{\theta}_{p_{\mathcal{U}_v}}(s_v)$ is the estimator of $\theta_{\mathcal{U}_v}$ based on the sample s_v. Similarly, the *p-variance* of the estimator $\hat{\theta}_{p_{\mathcal{U}_v}}$ is defined as $\mathrm{Var}(\hat{\theta}_{p_{\mathcal{U}_v}}) = E_{p_{\mathcal{U}_v}}(\hat{\theta}_{p_{\mathcal{U}_v}} - E_{p_{\mathcal{U}_v}}(\hat{\theta}_{p_{\mathcal{U}_v}}))^2$.

- p-MSE – The *p-mean squared error* for $\hat{\theta}_{p_{\mathcal{U}_v}}$ is defined as $E_{p_{\mathcal{U}_v}}(\hat{\theta}_{p_{\mathcal{U}_v}} - \theta_{\mathcal{U}_v})^2$

- p-asymptotically unbiased – An estimator $\hat{\theta}_{p_{\mathcal{U}_v}}$ is *p-asymptotically unbiased* for $\theta_{\mathcal{U}_v}$ if $\lim_{v\to\infty}[E_{p_{\mathcal{U}_v}}(\hat{\theta}_{p_{\mathcal{U}_v}}) - \theta_{\mathcal{U}_v}] = 0$

- p-consistent – An estimator $\hat{\theta}_{p_{\mathcal{U}_v}}$ is *p-consistent* for $\theta_{\mathcal{U}_v}$ if $\lim_{v\to\infty} \Pr(|\hat{\theta}_{p_{\mathcal{U}_v}} - \theta_{\mathcal{U}_v}| > \epsilon) = 0$

- p-consistent for a finite population – An estimator $\hat{\theta}_{p\mathcal{U}_v}$ of $\theta_{\mathcal{U}_v}$ is *p-consistent for a finite population* under a given class of designs if the sample equals the population, i.e. $s_v = \mathcal{U}_v$ implies that $\hat{\theta}_{p\mathcal{U}_v}(s_v) = \theta_{\mathcal{U}_v}$.

The definitions regarding the asymptotics are still somewhat nebulous because the limit process has not been fully specified. For example, in a cluster sample design, as the size of the population grows, the number of sampled clusters can grow while the number of people within each cluster remain constant, or the number of clusters can remain constant as the number of people within each cluster grow.

To evaluate the model-based estimators, both the randomization (or p) and the model (or ξ) distributions need to be considered. Some examples that were used in the previous section include ξp unbiased ($E_\xi E_p(\hat{\theta}_p) = \theta$) and ξp variance ($\mathrm{Var}_\xi(E_p(\hat{\theta}|Y_C)) + E_\xi(\mathrm{Var}_p(\hat{\theta}|Y_C))$), see Sarndal (1978). These criteria are used when comparing and evaluating the different estimators in Chapter 2.

1.2.5 Weighting Controversies

Regardless of the model-based approach used, a controversy exists regarding the role of the design-based sampling weights in model-based analysis. This controversy is an extension of the controversy regarding the use of design- or model-based analysis. Design-based analysts are generally very concerned about violating model-based assuptions. In Little (2004), the design- versus model-based debate is referred to as "inferential schizophrenia". The first type of compromise between the model-based and design-based analyses was described in the hybrid approach of Section 1.2.3. The next step to a fully model-based approach is to incorporate the sampling design into the model in such a way that does not use the sampling weights.

As seen in the examples in Sections 1.2.1 and 1.2.2, the model-based estimators, when the design is incorporated into the analysis, do sometimes match the design-based estimators. The controversy over the weights revolves around *how* the sampling design should be

incorporated into the analysis, not around *whether* the sampling design should be incorporated into the analysis. When discussing the weighting controversy, the options are to use weights in the analysis in a design-based fashion, or to not use weights in the analysis and incorporate the relevant aspects of the sampling design explicitly into the model.

Researchers who use sampling weights believe that the weights add robustness to the model-based analysis, reducing the dependence on the model assumptions, see Kalton (1989), Thomas and Cyr (2002), Patterson et al. (2002), and Vermunt and Magidson (2007), to name a few. If the model does not hold for the census, then the weighted estimates are the best approximation of the model parameter under a given distance function (i.e. Kullbeck-Leibler distance for MLE, least squares distance for regression, etc), see Pfeffermann (1993). Furthermore, the details of the sampling design are not always available due to confidentiality reasons, making sufficient modeling difficult.

Researchers who do not add sampling weights believe that the weights themselves impose a model and assumptions that are not clearly explained. In addition, adding weights clouds interpretation and inflates variances, see Hoem (1989), Fienberg (1989), Mislevy and Sheehan (1989b). As an early example regarding non-linear models, Hoem (1989) and Kalton (1989) debate over the use of weights in a Markov chain analysis.

Both design- and model-based researchers agree that sampling weights may help reduce the effects of informative sampling. However, recall that NSAR sampling is defined with respect to a given set of covariates. One way to reduce the effect of informative sampling is to include as many of the variables that define the sampling design into the covariate set, as possible.

Some researchers advocate the rule of thumb that if the weighted and unweighted analyses are statistically the same, then there is no informative sampling and the results are valid, see Pfeffermann (1993). This leads to many researchers running both weighted and unweighted analyses, and if they do not match the results go into the file drawer, see Lohr and Liu (1994) for example.

In addition to the confusion over design- and model-based analysis, there is no consensus on how weights should be incorporated into model based analyses. Pfeffermann (1993) presents six options (all with good literature reviews) for including weights in analyses:

1. Modifications of model dependent estimators or model-assisted survey sampling (Sarndal et al., 1992)
2. Restriction to models that yield design consistent estimators (Little, 1983)
3. Pseudo-likelihood approach (Skinner, 1989a)
4. Weighted estimating functions (Godambe and Thompson, 1986)
5. Use of sampling weights as surrogates of the design variables (Rubin, 1985)
6. MLE's derived from weighted distributions (Pfeffermann et al., 2006).

Items 1, 3, 4 and 6 above are all variations of the hybrid method from Section 1.2.3, with some being model-based and some being design-based. Many of the different methods provide similar estimators for a given model, but they do not all match and there is no consensus as to which method of inserting weights is best.

1.3 Linear Models

Next consider a multiple regression model. This analysis will be expanded upon in Chapter 2 when linear mixed-effects models are analyzed. The regression model in the next section are is an example of a model-based analysis that can incorporate the stratified sampling design.

The structure of linear models make them a natural choice for modeling the sampling design in surveys. This section first describes the basics of multiple regression, which can incorporate stratification from the sampling design, but not clustering. Next, design-based analysis of regression models are explained. Finally, linear mixed-effects models are introduced and their role with surveys discussed.

1.3.1 Model-Based Multiple Regression

Multiple regression models estimate the linear relationship between a variable of interest, Y with a set of covariates X,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon,$$

where $\epsilon_i \sim N(0, \sigma^2)$. Consider an extension of the mean of the population example from Sections 1.2.1, 1.2.2 and 1.2.3. Suppose the outcome variable Y is related to a covariate X and we want to compute the value of Y given we know the values of X. For example, let the outcome variable Y be purchase price of an individual's last car and let the covariate X be their income. Let there be H additional indicator variable covariates, one for each of the H strata. The model becomes,

$$y_{hi} = \beta_1 I_{h=1} + \beta_2 I_{h=2} + \cdots + \beta_H I_{h=H} + \beta_{H+1} x_{hi1} + \epsilon_{hi},$$

where $I_{h=1}$ is one if $h = 1$ and zero otherwise. Notice that the intercept has been removed from the model for identifiability. The error structure can be such that each stratum has its own error variance, or they can have a common error variance. Using standard techniques, for example see Weisberg (2005), the β coefficients (assuming common variance) can be estimated as

$$\hat{\beta} = (X^T X)^{-1}(X^T Y) \tag{1.3}$$

where the X matrix contains all the covariates in the model. The estimate of y_{hi} is then

$$\hat{y}_{hi} = \hat{\beta}_1 I_{h=1} + \hat{\beta}_2 I_{h=2} + \cdots + \hat{\beta}_H I_{h=H} + \hat{\beta}_{H+1} x_{hi1},$$

To get an estimate of the population mean, each non-sampled element can be estimated using the regression equation if the x_{hi1} covariate is known for each element in the population. If it is not known for the non-sampled estimates, then an estimate of the income for the non-sampled estimates can be used, for example the mean income for the given strata.

Suppose that the effect of the covariate x varies according to strata. The estimate of $\hat{\beta}_{H+1}$ from the above regression is a weighted estimate of the slope variances according to the proportion of elements in each stratum in the *sample.* The model can be updated to include interactions of stratum level indicators to estimate the effect for each stratum separately,

$$\begin{aligned} y_{hi} &= \beta_1 I_{h=1} + \beta_2 I_{h=2} + \cdots + \beta_H I_{h=H} \\ &+ \beta_{H+1} I_{h=1} x_{hi1} + \beta_{H+2} I_{h=2} x_{hi1} + \cdots + \beta_{H+H} I_{h=H} x_{hi1} + \epsilon_{hi}. \end{aligned}$$

As can be seen, this model contains many covariates even with this simple stratified only sampling design.

1.3.2 Survey Based Regression Models

The same example can be analyzed using design-based methods. For the design-based method, the Y variable, for example the purchase price of an individual's last car, is considered fixed. The goal is to find the least squares estimate of β_0 and β_1 from the model

$$y_{hi} = \beta_0 + \beta_1 x_{hi}$$

assuming that a census was taken. Note that this is not assumed to be a generating model, and there is no random error (the data are fixed). The parameters β_0 and β_1 are considered descriptive summaries of the population. The standard way to obtain this estimate is to use a sample-weighted least-squares estimator. That is, to obtain the estimates of β_0 and

β_1 that minimize

$$\sum_{h=1}^{H}\sum_{i=1}^{N_h} w_{hi} I_{hi}(y_{hi} - \beta_0 - \beta_1 x_{hi1})^2$$

where w_{hi} are the inverse probability weights. Note this is design unbiased (or p-unbiased from Section 1.2.4) for the census sum of squares,

$$\begin{aligned} E_p(\sum_{h=1}^{H}\sum_{i=1}^{N_h} w_{hi} I_{hi}(y_{hi} - \beta_0 - \beta_1 x_{hi1})^2) &= \sum_{h=1}^{H}\sum_{i=1}^{N_h} w_{hi} E_p(I_{hi})(y_{hi} - \beta_0 - \beta_1 x_{hi1})^2 \\ &= \sum_{h=1}^{H}\sum_{i=1}^{N_h} w_{hi}\frac{1}{w_{hi}}(y_{hi} - \beta_0 - \beta_1 x_{hi1})^2. \end{aligned}$$

This leads to a design-based estimator (Korn and Graubard, 1999) of

$$\hat{\beta}_w = (X^T W X)^{-1} X^T W Y,$$

where W is a diagonal matrix of sampling weights. Lohr (1999) provides an alternate derivation based on a hybrid method.

In survey based regression models, the effects of stratification and clustering are incorporated using the sampling weights instead of inserting strata indicator variables, as was done in the previous section. The only covariate in the survey regression model is x_{hi1}, for example, the individual's income level. The estimates from a survey regression analysis estimate β_0 and β_1, summary statistics on the census data. The proportions of elements within strata for these estimates match the proportions of elements in the finite population.

While this has the same form as the classic weighted-regression estimator, see Weisberg (2005), the classic weighted regression is used when there is unequal error variance. The weights used for unequal error variance are proportional to the inverse of the error variance. So, in classic weighted regression, the larger the weight, the smaller the error variance, the *more* is known about the data. With sampling weights, the weight represents the number

of elements in the population that the sampled person represents. Thus, the higher the weight the *less* is known about the data. Estimates of variance are appropriately modified.

This example demonstrates that for a multiple regression model, there are differences between model- and design-based analyses. As the models become more complex, so do the differences between the estimators. (Lohr, 1999, Chapter 11) contains a summary of the different regression parameter estimates along with a section on whether weights should be used in the regression analysis.

1.3.3 Linear Mixed-Effects (LME) Models

Model-based multiple regression models incorporate the strata effects, however they do not incorporate the clustering effects. To incorporate the clustering effects, random-effects need to be added. Linear mixed-effects models incorporate both fixed and random effects. A detailed description of linear mixed-effects models is in Chapter 2. To provide an overview, we consider a model that can incorporate stratification and clustering, a random intercept model. Suppose the sampling design is to cluster and then stratify. Let k represent clusters and h represent strata. Then the model

$$y_{khi} = \beta_1 I_{h=1} + \cdots + \beta_H I_{h=H} + U_{0k} + \epsilon_{hik}$$
$$U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon_{hik} \sim N(0, \sigma^2_{\epsilon}), \quad U_{0k} \perp \epsilon$$

corresponds to a clustered then stratified sampling design, provided the strata are individually labeled so the stratum indicator implies appropriate cluster inclusion. The random intercept, U_{0k} varies according to cluster. Two elements in the same cluster, regardless of strata membership, have a covariance of σ^2_{0k}. The covariance between two elements in different clusters is zero. The variance of an individual element is $\sigma^2_{0k} + \sigma^2_{\epsilon}$. The modeling of the variance components in linear mixed-effects models can match very complex sampling designs. Inserting sampling weights into linear mixed-effects models and a comparison

of the weighted and unweighted analyses under model misspecification and informative sampling are examined in depth in Chapters 2 and 3.

1.4 Thesis Outline

The goal of this dissertation is to investigate the use of design-based sampling weights in model-based analysis. The model-based analyses include frequentist linear mixed effects models and a Bayesian grade of membership model.

Chapter 2 investigates the theory of mixed-effects models with complex survey sampling data. There are three published approaches for inserting weights in mixed-effects models, Pfeffermann et al. (1998), Korn and Graubard (2003) and Rabe-Hesketh and Skrondal (2006), the latter published concurrently by Asparouhov (2006). Chapter 2 derives the approaches from a common starting point and shows where decisions are made that differentiate the approaches.

Chapter 3 of this thesis provides a simulation study comparing weighting methods in LME models. These simulations evaluate the benefits and drawbacks of using the weights in LME models under different levels of model misspecification and informative sampling.

Chapter 4 extends what was learned in chapters 2 and 3 to Bayesian grade of membership models. Prior distributions are changed to incorporate sampling design in a model-based analysis. Sampling weights are also incorporated into the analysis. Simulation runs compare the effect of informative sampling on both unweighted and weighted analyses.

Chapter 5 summarizes the findings and provides recommendations for future work.

1.5 Thesis Contributions

This thesis clarifies some of the confusion regarding the use of sampling weights in model-based analyses. By providing side-by-side comparisons of different methods, in theory and in practice, decisions can be made regarding the effects of sampling weights. The

contributions of this thesis include, but are not limited to, the following.

In Chapter 2, three approaches for inserting weights into linear mixed-effects models are derived from a common starting point and decisions that differentiate the approaches are highlighted. This demonstrates the ad-hoc nature of the pseudo-likelihood technique as the different approaches insert sampling weights at different locations in the estimation process. After deriving the approaches, I prove necessary conditions for consistency. I review three different published scalings of the sample weights and derive analytically the differences of the scalings in a random intercept model. I provide bounds for the bias of the parameter estimates under different scaling methods. Finally, I derive conditions under which two of the methods equal each other.

In Chapter 3, I execute an extensive simulation study for linear mixed-effects models that demonstrates the effects of model misspecification and informative sampling on the different weighted estimates. These simulations show that the two major approaches from Chapter 2 provide very similar estimates in practice. The sandwich estimator for the variance outperforms the design-based approximation for the variance. Sampling weights do not help model misspecification unless it induces informative sampling. Sampling weights do help protect against informative sampling, but do not correct it entirely. I created initial metrics to compare the estimates from the models across all parameters. From these, I found that the unweighted analyses performed very well under model misspecification when there is no informative sampling due to their reduced variance. The unweighted estimates also performed well under informative sampling, however that is dependent on the magnitude of informative sampling in the simulations. The weighted estimates using the scaled 2 weights (see Section 2.4.2) performed well when there was a significant amount of informative sampling. In addition, I created a new graphic to easily compare multiple simulations on one page.

In Chapter 4, I propose a modified version of the GoM model to incorporate the dependencies induced by the sampling design. This modification incorporates a polytomous

logistic mixed-effects regression analysis, which also can be used to incorporate dependencies due to longitudinal data. I next propose a principled way of weighting the GoM analysis, called weighting based on the estimated parameter. This weighting is an extension of the PML weighting of Chapter 2, but also weights depending on the type of parameter being estimated. Finally, I execute a simulation study to evaluate the performance of the polytomous logistic mixed-effects regression prior in both unweighted and weighted based on the estimated parameter estimates. These simulations demonstrate that the effect of the sampling design and the weights is generally similar to what was found in Chapter 3.

In Chapter 5, I provide direction for future research.

Chapter 2

Sampling Weights in Linear Mixed-Effects Models

Linear mixed-effects (LME) models analyze data that contain complex patterns of variability, specifically involving different nested layers. Examples include longitudinal data (multiple observations per person), educational data (multiple pupils per classroom), survey data (units within clusters within strata), etc. While LME models can match well the stratification and clustering of survey data, the debate continues whether or not sampling weights should be used and, if used, how they should be incorporated into LME estimates. This chapter compares proposals on how to insert sampling weights into LME models and Chapter 3 compares the proposals using simulated data.

Section 2.1 summarizes the basic structure of the LME model using a model-based frequentist framework. Section 2.2 introduces pseudo-maximum likelihood and reviews three current published proposals for using sample weights in obtaining point estimates of LME parameters. Section 2.3 reviews the distinct approaches each proposal implements to obtain variances of the point estimates. In Section 2.4, I evaluate the different methods with respect to consistency. The section continues by describing three different scalings of the weights that are used by these authors. I conclude this section by comparing the

methods from Rabe-Hesketh and Skrondal (2006) and Pfeffermann et al. (1998) for the random intercept case, and I state conditions upon which the two methods provide the same result.

In addition to summarizing relevant information in one place, my contributions in this chapter include the use of a common notation for all methods and the comparison of all methods from a common starting point so I can emphasize where each method makes unique decisions (such as when the sampling weights are inserted). In addition, my description of Pfeffermann et al. (1998) explains the origins of the algebraic manipulations that are obscured by their cryptic notations. These contributions facilitate the understanding of and the comparisons between methods.

I add to the content of the research by providing detailed derivation of conditions necessary for consistency. I develop theoretical bounds on the variance components in the random intercept model when using the three different scalings of the weights. These bounds allow a theoretical comparison between methods and are referred to when comparing simulation results in Chapter 3. Finally, I provide conditions upon which Rabe-Hesketh and Skrondal (2006) and Pfeffermann et al. (1998) provide identical estimates for a random intercept case.

2.1 Linear Mixed-Effects Models

Many national surveys utilize complex multi-stage sampling designs that induce complex variance components. Linear mixed-effects (LME) models, developed in Laird and Ware (1982) and described in depth in Searle et al. (1992), analyze data with such groupings. The model, described in the next section, is

$$Y = X\beta + ZU + \epsilon, \quad U \sim N(0, \Omega), \epsilon \sim N(0, \sigma_\epsilon^2 I),$$

where Y is a vector of response variables, X is the matrix of fixed effect variables, β is the vector of fixed effect coefficients, Z is the matrix of random effect variables, U is the vector of random effect coefficients and ϵ is a vector of random errors. This model incorporates cluster and stratification sampling layers in the X and Z matrices. Examples of surveys using multi-stage sampling include the National Survey of Child and Adolescent Well-Being (Down et al., 2002), the National Assessment of Educational Progress (Vinovskis, 1998), and the National Long Term Care Survey (NLTCS, 1988).

As outlined in Section 1.2.5, some researchers always use sampling weights when estimating model parameters. Incorporating sampling weights in LME parameter estimation is not straightforward. For example, Pfeffermann et al. (1998), Korn and Graubard (2003), and Rabe-Hesketh and Skrondal (2006) document three competing approaches in which the ad-hoc Horvitz-Thompson approach described in Section 1.2.1 can be used to incorporate the sampling weights into LME models. This chapter begins by reviewing unweighted LME model theory, and then compares and contrasts these three approaches on adding sampling weights to LME models. Chapter 3 compares the different approaches using simulation studies.

2.1.1 Notation and Parameterization of Random Effects

Let Y $(N \times 1)$, Y_k $(N_k \times 1)$ and Y_{ik} represent the outcome variable for the entire data set, for cluster k and for individual ik, respectively, where N_k is the number of elements sampled from cluster k and $\sum_{k=1}^{K} N_k = N$. For simplicity, assume one level of clustering. The Y vector stacks the outcomes from each cluster,

$$Y = \begin{bmatrix} Y_{11} & Y_{21} & \cdots & Y_{N_1 1} & \cdots & Y_{1K} & Y_{2K} & \cdots & Y_{N_K K} \end{bmatrix}^T = \begin{bmatrix} Y_1^T & Y_2^T & \cdots & Y_K^T \end{bmatrix}^T,$$

where K is the total number of clusters. Let X $(N \times P)$, β $(P \times 1)$, represent the matrix of fixed effect variables and fixed effect coefficients, respectively. X can be partitioned into

K sub-matrices, each representing a cluster. Let x_{ipk} be the value of the p^{th} variable for the i^{th} element in cluster k. Then

$$X\beta = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{bmatrix} \beta \quad \text{where} \quad X_k = \begin{bmatrix} x_{11k} & x_{12k} & \cdots & x_{1Pk} \\ x_{21k} & x_{22k} & \cdots & x_{2Pk} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_k1k} & x_{N_k2k} & \cdots & x_{N_kPk} \end{bmatrix}.$$

Let Z $(N \times KQ)$ represent the matrix of random effect variables, where Q is the number of random effects. Let U $(KQ \times 1)$ represent the vector of random effect coefficients. Let Z_q (U_q) $(N \times K, K \times 1)$, represent the matrix of random effect variables (coefficients) containing the values of the the q^{th} variable (coefficient) for each of the K clusters, respectively. Then

$$ZU = \begin{bmatrix} Z_1 & Z_2 & \cdots & Z_Q \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_Q \end{bmatrix} = Z_1U_1 + Z_2U_2 + \cdots + Z_QU_Q.$$

Z and U are comprised of Z_q and U_q as shown in Equation 2.1 below, where Z_{iqk} represents the q^{th} random effect variable for the i^{th} element in cluster k, and U_{qk} represents the q^{th} random effect coefficient for the k^{th} cluster. Let Z_{qk} $(N_k \times 1)$ represent a vector of

the q^{th} random effect variable for cluster k. Then

$$Z_q U_q = \begin{bmatrix} Z_{1q1} & 0 & \cdots & 0 \\ Z_{2q1} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ Z_{N_1 q1} & 0 & \cdots & 0 \\ 0 & Z_{1q2} & \cdots & 0 \\ 0 & Z_{2q2} & \cdots & 0 \\ & \vdots & \ddots & 0 \\ 0 & Z_{N_2 q2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & Z_{1qK} \\ 0 & 0 & \cdots & Z_{2qK} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_{N_K qK} \end{bmatrix} \begin{bmatrix} U_{q1} \\ U_{q2} \\ \vdots \\ U_{qK} \end{bmatrix} = \begin{bmatrix} Z_{q1} & 0 & \cdots & 0 \\ 0 & Z_{q2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_{qK} \end{bmatrix} \begin{bmatrix} U_{q1} \\ U_{q2} \\ \vdots \\ U_{qK} \end{bmatrix} \quad (2.1)$$

The multiple regression models of Section 1.3.1 incorporate the fixed effect variables, but not the random effect variables.

2.1.2 Model-Based Frequentist Linear Mixed-Effects Models

Laird and Ware (1982) outlined LME models for data with variance components induced by groupings. Equation 2.2 below contains a two-level model, the generalization to a multi-level LME model is described in Snijders and Bosker (1999). Let

$$\left.\begin{array}{rcc} Y & = & X\beta + ZU + \epsilon \\ & = & X\beta + \sum_{q=1}^{Q} Z_q U_q + \epsilon \\ & \text{where} & U_q \sim N(0, \omega_{qq}^2 I_{K\times K}), \epsilon \sim N(0, \sigma_e^2 I_{N\times N}), \\ & & \text{cov}(U_q, U_r) = \omega_{qr}^2 I_{K\times K}, U \perp \epsilon \end{array}\right\}. \quad (2.2)$$

The X, β, Z and U matrices are of the form described previously. Let ϵ $(N \times 1)$ represent the random error. Consider a model with a random intercept and a random slope. Then $\text{Var}(U)$ can be partitioned into four sub-matrices,

$$\text{Var}(U) = \begin{bmatrix} \text{Var}(U_1) & \text{Var}(U_1, U_2) \\ \text{Var}(U_2, U_1) & \text{Var}(U_2) \end{bmatrix} = \Omega.$$

The submatrices $\text{Var}(U_1) = \omega_{11}^2 I_{K\times K}$ and $\text{Var}(U_2) = \omega_{22}^2 I_{K\times K}$ contain the variances for the random intercept (slope) for each cluster on the diagonal, and zero elements on the off-diagonal indicating independence between the intercept (slope) for cluster k_1 and the intercept (slope) from cluster k_2, respectively. The submatrices $\text{Var}(U_1, U_2) = \text{Var}(U_2, U_1) = w_{12}^2 I_{K\times K}$ are diagonal matrices where the diagonal element is the covariance between the intercept and the slope within each cluster. The off-diagonal zero elements indicate that the intercept (slope) from cluster k_1 is independent of the slope (intercept) of cluster k_2. If there are more than two random effects, the $\text{Var}(U)$ matrix has the form

$$\text{Var}(U) = \Omega = \begin{bmatrix} \omega_{11} I_{KxK} & \omega_{12} I_{KxK} & \cdots & \omega_{1Q} I_{KxK} \\ \omega_{21} I_{KxK} & \omega_{22} I_{KxK} & \cdots & \omega_{2Q} I_{KxK} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{Q1} I_{KxK} & \omega_{Q2} I_{KxK} & \cdots & \omega_{QQ} I_{KxK} \end{bmatrix} = \mathcal{O} \otimes I_{K\times K} \tag{2.3}$$

where $\otimes$ represents the Kronecker product and $\mathcal{O}$ is the matrix of the $\omega_{q_1 q_2}$'s. Given the model in Equation 2.2, the variance of Y can be computed as

$$V = \text{Var}(Y) = \text{Var}(XB + ZU + \epsilon) = Z\Omega Z^T + \sigma_\epsilon^2 I_{N\times N}. \tag{2.4}$$

It is easily shown that $V^{-1}V = I_{N\times N}$ when V^{-1} is defined as

$$V^{-1} = \sigma_\epsilon^{-2}\left(I_{N\times N} - ZAZ^T\right) \tag{2.5}$$

where

$$A = \left(Z^T Z + \sigma_\epsilon^2 \Omega^{-1}\right)^{-1}. \quad (2.6)$$

This parameterization is described in depth in Searle et al. (1992) and Littell et al. (1996).

2.2 Methods for Weighting LME Models – Point Estimates

To ease notation throughout this chapter, we denote the three approaches that incorporate the sampling weights into LME models as follows, PSHGR for Pfeffermann et al. (1998), KG for Korn and Graubard (2003), and RHS for Rabe-Hesketh and Skrondal (2006). In Section 2.3 we describe how the PSHGR, KG and RHS approaches compute standard errors of estimation. Finally, in Section 2.4 we compare these approaches.

Recall from Section 1.2.1 that the sampling weight definition is $w_{ki} = \frac{1}{\pi_{ki}}$, where π_{ki} is the probability that element ki is included in the sample. The sampling weight, w_{ki} can be interpreted as the number non-sampled elements in the finite population that sampled element ki represents. The approaches in this chapter use various conditionings of the sampling weights. Suppose the sampling design first clusters and then samples elements within a cluster. Let $\pi_{ki} = \pi_k \pi_{i|k}$, where π_k is the probability that cluster k is sampled and $\pi_{i|k}$ is the probability that element i in cluster k is sampled. The corresponding weights are defined as $w_k = \frac{1}{\pi_k}$ and $w_{i|k} = \frac{1}{\pi_{i|k}}$. Higher order conditional weights are defined similarly. For example, let π_{ijk} be the probability that elements i and j in cluster k are sampled. Let $\pi_{ijk} = \pi_k \pi_{ij|k}$ where $\pi_{ij|k}$ is the probability that both elements i and j are sampled provided cluster k is sampled. The corresponding weights are $w_k = \frac{1}{\pi_k}$ and $w_{ij|k} = \frac{1}{\pi_{ij|k}}$. The approaches in this chapter use the following weights: cluster weights, w_k, (univariate) conditional weights $w_{i|k}$, bvariate conditional weights $w_{ij|k}$, trivariate conditional weights $w_{lms|k}$ and quadvariate conditional weights, $w_{lmst|k}$.

2.2.1 Maximum Likelihood Estimation

Maximum likelihood (ML) estimates have many favorable properties, such as asymptotic efficiency, asymptotic consistency, and asymptotic normality. Survey practitioners want to capture the good properties of ML estimates while incorporating the survey design into the analysis. Chambless and Boyle (1985) provide conditions upon which a pseudo-maximum likelihood estimate for a stratified sample is asymptotically unbiased and asympotically Normal. Binder (1983) provides a general way to estimate the covariance matrix. This subsection computes the likelihood and score functions of the LME model, and the next subsection describes approaches to incorporating the sampling weights.

From Equation 2.2, the likelihood is

$$L(Y|X,Z,\beta,\Omega,\sigma_\epsilon^2) = \frac{\exp\left(-\frac{1}{2}(Y-X\beta)^T V^{-1}(Y-X\beta)\right)}{(2\pi)^{\frac{N}{2}}|V|^{\frac{1}{2}}}, \tag{2.7}$$

where $V = \mathrm{Var}(Y)$ from Equation 2.4. The log likelihood is

$$\begin{aligned} l &= \log L(Y|X,Z,\beta,\Omega,\sigma_\epsilon^2) \\ &= -\frac{1}{2}N\log(2\pi) - \frac{1}{2}\log|V| - \frac{1}{2}(Y-XB)^T V^{-1}(Y-XB), \end{aligned} \tag{2.8}$$

which can be maximized by setting the score functions

$$l_\beta = \frac{\partial l}{\partial \beta} = X^T V^{-1} Y - X^T V^{-1} X\beta \tag{2.9}$$

$$\begin{aligned} l_{\omega_{ij}} = \frac{\partial l}{\partial \omega_{ij}} &= -\frac{1}{2}\mathrm{tr}\left(V^{-1}(Z_i Z_j^T + \delta_{i\neq j} Z_j Z_i^T\right) \\ &+ \frac{1}{2}(Y-X\beta)^T V^{-1}\left(Z_i Z_j^T + \delta_{i\neq j} Z_j Z_i^T\right) V^{-1}(Y-X\beta) \end{aligned} \tag{2.10}$$

equal to zero and solving for the parameter estimates. In this parameterization, the derivative with respect to σ_ϵ^2 can be obtained by letting $Z_i = Z_j = I_{N\times N}$ in Equation 2.10. Details of the derivative calculation can be found in Section 6.2 of Searle et al. (1992).

2.2.2 Pseudo-Maximum Likelihood (PML) Estimation

The PML approach estimates the census likelihood equations in Equations 2.7 to 2.10 with the weighted sample data. From Section 1.2.5, using the sample without incorporating the sampling design may bias the results. Recall a sampling weight approximates the number of people in the census that the sampled element represents. The PML approaches take the sample likelihood, and replicate each sampled element by the value of its weight to estimate the census likelihood. The parameter are estimated using the weighted sample likelihood. This provides accurate point estimates, but not accurate variances, as described further in Section 2.3.

The three approaches differ in how the weights are used to replicate the sampled elements. RHS insert the weights before the derivative is taken, around Equation 2.7. KG insert the weights immediately after the derivative is taken, around Equation 2.9 and 2.10. PSHGR insert the weights in the process of solving for the parameter values using Equations 2.9 and 2.10.

2.2.3 RHS Weighting

RHS estimate the census likelihood by weighting the sample likelihood. This weighted sample likelihood can be maximized using any technique, and the stata function `gllamm()` uses the Newton-Rhapson method.

To create this pseudo-likelihood, we begin by writing the census likelihood as

$$\begin{aligned} L(Y|X,Z,\beta,\Omega,\sigma_\epsilon^2) &= \prod_{k=1}^{K} L(Y_k|X,Z,\beta,\Omega,\sigma_\epsilon^2) \\ &= \prod_{k=1}^{K} \int_{U_k} \left[\prod_{i=1}^{N_k} L(Y_{ik}|U_k,X,Z,\Omega,\beta,\sigma_\epsilon^2) \right] L(U_k|\Omega) dU_k. \end{aligned}$$

This census likelihood conditions on the value of the random effects, U_k's, to create independence within a cluster. These U_k values are integrated out to get a likelihood that is

not conditional upon the random effects. To estimate this with sample data, RHS rewrite the census likelihood as a likelihood over sampled elements, and replicate the sampled Y_{ik}'s by their weight, $w_{i|k}$,

$$L_w(Y_k|X,Z,\beta,\Omega,\sigma_\epsilon^2) = \int_{U_k}\left[\prod_{i=1}^{n_k} L(Y_{ik}|U_k,X,Z,\beta,\Omega,\sigma_\epsilon^2)^{w_{i|k}}\right] L(U_k|\Omega)dU_k \quad (2.11)$$

where $w_{i|k}$ is the inverse probability of sampling individual i given cluster k is sampled. To estimate $L(Y|X,Z,\beta,\Omega,\sigma_\epsilon^2)$, RHS also replicate the clusters according their weights, w_k,

$$\begin{aligned} L_w(Y|X,Z,\beta,\Omega,\sigma_\epsilon^2) &= \prod_{k=1}^{k_s} L_w(Y_k|X,Z,\beta,\Omega,\sigma_\epsilon^2)^{w_k} \\ &= \prod_{k=1}^{k_s}\left(\int_{U_k}\left[\prod_{i=1}^{n_k} L(Y_{ik}|U_k,X,Z,\beta,\Omega,\sigma_\epsilon^2)^{w_{i|k}}\right] L(U_k|\Omega)dU_k\right)^{w_k} \quad (2.12) \end{aligned}$$

where w_k is the inverse probability that cluster k is included in the sample and k_s is the number of sampled clusters. Inserting the normal density for a two level model into Equations 2.11 and 2.12 provides the weighted sample likelihood,

$$\begin{aligned} L_w(Y_k|X,Z,\beta,\Omega,\sigma_\epsilon^2) &= \int_{U_k}\left[\prod_{i=1}^{n_k} L(Y_{ik}|U_k,X,Z,\beta,\Omega,\sigma_\epsilon^2)^{w_{i|k}}\right] L(U_k|\Omega)dU_k \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_\epsilon}\sum_{i=1}^{n_k} w_{i|k}(y_{ik}-x_{ik}\beta)^2\right)\right\} \\ &\times \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n_k} w_{i|k}\left\{(y_{ik}-x_{ik}\beta)^T Z_{ik}^T\right\}[A_{kw}]\sum_{i=1}^{n_k} w_{i|k}\left\{Z_{ik}(y_{ik}-x_{ik}\beta)\right\}\right)\right\} \\ &= \exp\left(-\frac{1}{2}\left\{(Y_k-X_k\beta)^T\sigma_\epsilon^{-2}\left[W_{\cdot|k}-W_{\cdot|k}Z_kA_{kw}Z_k^TW_{\cdot|k}\right](Y_k-X_k\beta)\right\}\right) \end{aligned}$$

where $A_{kw} = (Z_k^TW_{\cdot|k}Z_k + \sigma_\epsilon^2\Omega^{-1})^{-1}$, and $W_{\cdot|k}$ is a $n_k \times n_k$ diagonal matrix with the $w_{i|k}$ weights for the cluster on the diagonal. From this,

$$L_w(Y_k|X,Z,\beta,\Omega,\sigma_\epsilon^2) \sim N(X_k\beta, \sigma_\epsilon^{-2}[W_{\cdot|k}-W_{\cdot|k}Z_kA_{kw}Z_k^TW_{\cdot|k}]).$$

The $L_w(Y|X,Z,\beta,\Omega,\sigma_\epsilon^2)$ is obtained as

$$\begin{aligned} L_w(Y|X,Z,\beta,\Omega,\sigma_\epsilon^2) &= \prod_k^{k_s} L_w(Y_k|X,Z,\beta,\Omega,\sigma_\epsilon^2)^{w_k} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^{\hat{N}} \prod_{k=1}^{k_s} \left(\sigma_\epsilon^{-2}\left|W_{\cdot|k} - W_{\cdot|k} Z_k A_{kw} Z_k^T W_{\cdot|k}\right|\right)^{w_k} \\ &\times \exp\left(-\frac{1}{2}\sum_k^{k_s} w_k \left\{(Y_k - X_k\beta)^T \sigma_\epsilon^{-2}\left[W_{\cdot|k} - W_{\cdot|k} Z_k A_{kw} Z_k^T W_{\cdot|k}\right](Y_k - X_k\beta)\right\}\right) \end{aligned} \quad (2.13)$$

where $\hat{N} = \sum_{k=1}^{k_s}\sum_{i=1}^{n_k} w_k w_{i|k}$. This weighted log likelihood can be maximized to get weighted parameter estimates.

Note that Asparouhov (2006), denoted ASP in Chapter 3, published the same method the same year. I refer to this method as the RHS method as the software I used for the simulations in Chapter 3 was provided by RHS.

2.2.4 KG Weighting

KG weight the sample score function to estimate census score function in Equations 2.9 and 2.10. The weighted parameter estimates are obtained from the weighted score function.

KG rewrite the census score function from Equations 2.9 and 2.10 in terms of sums. Let $V_k^{-1} = (\text{var}(Y_k))^{-1}$, and let v_{ijk}^{-1} represent the ij^{th} element of V_k^{-1}. Then

$$l_\beta(V,\beta|y) = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k} x_{ipk} v_{ijk}^{-1}(Y - X\beta)_{jk} \quad p = 1,...,P \quad (2.14)$$

$$\begin{aligned} l(V,\beta|y) &= \sum_{k=1}^{K}\sum_{l=1}^{N_k}\sum_{m=1}^{N_k}\sum_{s=1}^{N_k}\sum_{t=1}^{N_k}(Y - X\beta)_{lk}(Y - X\beta)_{sk} v_{lmk}^{-1} v_{tsk}^{-1} z_{mik} z_{tjk} \\ &- \sum_{k=1}^{K}\sum_{s=1}^{N_k}\sum_{t=1}^{N_k} v_{stk}^{-1}(z_{sik} z_{tjk} + \delta_{s\neq t} z_{tik} z_{sjk}), \\ &\quad i = 1,...,Q, j = 1,...,Q \end{aligned} \quad (2.15)$$

$$l_{\sigma_\epsilon^2}(V,\beta|y) = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{N_k}\sum_{s=1}^{N_k}(Y - X\beta)_{ik} v_{ijk}^{-1} v_{jsk}^{-1}(Y - X\beta)_{sk} - \sum_{k=1}^{K}\sum_{i=1}^{N_k} v_{iik}^{-1}. \quad (2.16)$$

The above sums over the census quantities are estimated using weighted sums over sample quantities using the Horvitz-Thompson heuristic. Notice that Equations 2.14 , 2.15 and

2.16 contain one sum over clusters (indexed by k) and either one, two, three or four sums over individuals (indexed by i, j, l, m, s and t). To accommodate this, KG need the univariate, bivariate, trivariate and quadvariate conditional weights ($w_{i|k}, w_{ij|k}, w_{ijs|k}$ and $w_{lmst|k}$), which are the inverse probability of having the subscripted one, two, three or four elements in the sample given the cluster is sampled. The weighted sample score functions are

$$l^w_\beta(\beta,\Omega,\sigma^2_\epsilon|y) = \sum_{k=1}^{K} w_k I_k \sum_{i=1}^{N_k}\sum_{j=1}^{N_k} w_{ij|k} I_{ij|k} x_{ipk} v^{-1}_{ijk}(Y - X\beta)_{jk} \quad p = 1, ..., P \tag{2.17}$$

$$\begin{aligned} l^w_{\omega_{ij}}(\beta,\Omega,\sigma^2_\epsilon|y) &= \sum_{k=1}^{K} w_k I_k \sum_{l=1}^{N_k}\sum_{m=1}^{N_k}\sum_{s=1}^{N_k}\sum_{t=1}^{N_k} w_{lmst|k} I_{lmst|k}(Y - X\beta)_{lk}(Y - X\beta)_{sk} v^{-1}_{lmk} v^{-1}_{tsk} z_{mik} z_{tjk} \\ &- \sum_{k=1}^{K} w_k I_k \sum_{s=1}^{N_k}\sum_{t=1}^{N_k} w_{st|k} I_{st|k} v^{-1}_{stk}(z_{sik} z_{tjk} + \delta_{s\neq t} z_{tik} z_{sjk}), \\ &\quad i = 1, ..., Q, j = 1, ..., Q \end{aligned} \tag{2.18}$$

$$\begin{aligned} l^w_{\sigma^2_\epsilon}(V,\beta|y) &= \sum_{k=1}^{K} w_k I_k \sum_{i=1}^{N_k}\sum_{j=1}^{N_k}\sum_{s=1}^{N_k} w_{ijs|k} I_{ijs|k}(Y - X\beta)_{ik} v^{-1}_{ijk} v^{-1}_{jsk}(Y - X\beta)_{sk} \\ &- \sum_{k=1}^{K} w_k I_k \sum_{i=1}^{N_k} w_{i|k} I_{i|k} v^{-1}_{iik}. \end{aligned} \tag{2.19}$$

Indicator variables (i.e. $I_{ij|k} = 1$ if i and j are in the sample, 0 otherwise), which zero out elements that have not been sampled, are incorporated into the sums. The weighted likelihood equations can be solved using any standard equation solver.

The derivation for a general mixed effects model for the KG approach requires univariate, bivariate, trivariate and quadvariate conditional weights. To demonstrate that only univariate and bivariate conditional weights are needed in some simplified models, consider the model used in KG's derivation in their paper, a random intercept model with no

covariates, i.e. $Y_{ik} = \beta_0 + U_{0k} + \epsilon_{ik}$. Here,

$$\begin{aligned} V_k &= \mathrm{Var}(Y_k) \\ &= \sigma_\epsilon^2 I_{N_k \times N_k} + \sigma_{0k}^2 J_{N_k \times N_k} \\ V_k^{-1} &= \sigma_\epsilon^{-2} \left(I_{N_k \times N_k} - \frac{\sigma_{0k}^2}{N_k \sigma_{0k}^2 + \sigma_\epsilon^2} J_{N_k \times N_k} \right) \\ v_{ij|k}^{-1} &= \begin{cases} \frac{1}{\sigma_\epsilon^2 (N_k \sigma_{0k}^2 + \sigma_\epsilon^2)} \left((N_k - 1)\, \sigma_{0k}^2 + \sigma_\epsilon^2 \right) & \text{if } i = j \\ \frac{-1}{\sigma_\epsilon^2 (N_k \sigma_{0k}^2 + \sigma_\epsilon^2)} \sigma_{0k}^2 & \text{if } i \neq j \end{cases} \end{aligned}$$

where $J_{N_k \times N_k}$ is a matrix of ones. With $x_{ipk} = 1$, Equations 2.14 to 2.16 become

$$l_\beta = \sum_{k=1}^{K} \sum_{j=1}^{N_k} \frac{Y_{ik}}{N_k \sigma_{0k}^2 + \sigma_\epsilon^2} - \beta \sum_{k=1}^{K} \frac{N_k}{N_k \sigma_{0k}^2 + \sigma_\epsilon^2} \tag{2.20}$$

$$l_{\sigma_{0k}^2} = \sum_{k=1}^{K} \sum_{l=1}^{N_k} \sum_{s=1}^{N_k} \frac{1}{N_k \sigma_{0k}^2 + \sigma_\epsilon^2} (Y - \beta)_{lk} (Y - \beta)_{sk} - \sum_{k=1}^{K} \frac{N_k}{N_k \sigma_{0k}^2 + \sigma_\epsilon^2} \tag{2.21}$$

$$\begin{aligned} l_{\sigma_\epsilon^2}^2 &= \frac{1}{\sigma_\epsilon^2} \sum_{k=1}^{K} \sum_{i=1}^{N_k} (Y - \beta)_{ik}^2 \left[1 - \frac{2\sigma_{0k}^2}{N_k \sigma_{0k}^2 + \sigma_\epsilon^2} + \frac{N_k \sigma_{0k}^4}{(N_k \sigma_{0k}^2 + \sigma_\epsilon^2)^2} \right] \\ &+ \frac{1}{\sigma_\epsilon^2} \sum_{k=1}^{K} \sum_{s=1}^{N_k} \sum_{i=1, i \neq s}^{N_k} (Y - \beta)_{ik} (Y - \beta)_{sk} \left[\frac{N_k \sigma_{0k}^4}{(N_k \sigma_{0k}^2 + \sigma_\epsilon^2)^2} - \frac{2\sigma_{0k}^2}{N_k \sigma_{0k}^2 + \sigma_\epsilon^2} \right] \\ &- \sum_{k=1}^{K} \frac{N_k}{\sigma_\epsilon^2} \left[1 - \frac{\sigma_{0k}^2}{N_k \sigma_{0k}^2 + \sigma_\epsilon^2} \right]. \end{aligned} \tag{2.22}$$

In Equations 2.20 to 2.22 the sums are over one or two individuals in a cluster. The univariate and bivariate conditional weights are sufficient for estimating these census equations.

The main drawback with this method for a general LME is obtaining the higher order (i.e. bivariate, triple and quadvariate) conditional weights. Most national surveys do not provide, or even calculate, this information.

2.2.5 PSHGR Weighting

PSHGR incorporate the sampling weights while solving Equations 2.9 and 2.10 for parameter estimates. As the census likelihood is a function of linear statistics, the estimated likelihood is a function of the weighted linear statistics. This carries independence assumptions as discussed in this section, which allows them to use univariate conditional weights, $w_{i|k}$, and not need higher order conditional weights. PSHGR use an iterative generalized least squares algorithm to estimate β and $\theta = \left[\text{vech}(\mathcal{O}) \;\; \sigma_\epsilon^2\right]$, where $\mathcal{O}$ is defined after Equation 2.3 and vech($\mathcal{O}$) contains the unique elements of $\mathcal{O}$. Writing the census likelihood equation as a function of linear statistics requires extensive matrix algebra manipulations. To demonstrate these, first consider the parameter estimates based on the census data.

Unweighted solution: $\hat{\beta}$

Because V is block diagonal by cluster, Equation 2.9 can be re-written as

$$0 = \sum_{k=0}^{K} X_k^T V_k^{-1} Y_k - \left(\sum_{k=0}^{K} X_k^T V_k^{-1} X_k \right) \beta$$

and β can be estimated as,

$$\begin{aligned} \hat{\beta} &= \left(\sum_{k=0}^{K} X_k^T V_k^{-1} X_k \right)^{-1} \left(\sum_{k=0}^{K} X_k^T V_k^{-1} Y_k \right) \\ &= P^{-1} T. \end{aligned} \tag{2.23}$$

Updates to P and T depend on the previous iteration of $\mathcal{O}$ estimates in the V^{-1} matrix.

Unweighted solution: $\hat{\theta}$, Random Intercept Case

Solving for θ (embedded in V^{-1}) directly from Equations 2.7 - 2.10 is difficult, and Goldstein (1986) provides an alternative method for computing the maximum likelihood estimates. First, consider a random intercept model, $Y_{ik} = X\beta + U_{0k} + \epsilon_{ik}, U_0 \sim N(0, \sigma_{0k}^2 I_{KxK}), \epsilon \sim$

$N(0, \sigma_\epsilon^2 I_{N\times N})$ with ϵ's independent of each other and U_{0k}. Let $\tilde{Y}_k = Y_k - X_k\beta$. By the variance definition, $E(\tilde{Y}_k\tilde{Y}_k^T) = \text{var}(Y_k) = V_k$. Then

$$E(\text{vec}(\tilde{Y}_k\tilde{Y}_k^T)) = E\begin{pmatrix} \tilde{Y}_{1k}\tilde{Y}_{1k} \\ \tilde{Y}_{1k}\tilde{Y}_{2k} \\ \vdots \\ \tilde{Y}_{1k}\tilde{Y}_{n_k k} \\ \vdots \\ \tilde{Y}_{n_k k}\tilde{Y}_{1k} \\ \tilde{Y}_{n_k k}\tilde{Y}_{2k} \\ \vdots \\ \tilde{Y}_{n_k k}\tilde{Y}_{n_k k} \end{pmatrix} = \begin{pmatrix} \sigma_{0k}^2 + \sigma_e^2 \\ \sigma_{0k}^2 \\ \vdots \\ \sigma_{0k}^2 \\ \vdots \\ \sigma_{0k}^2 \\ \sigma_{0k}^2 \\ \vdots \\ \sigma_{0k}^2 + \sigma_e^2 \end{pmatrix} = \sigma_{0k}^2 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \sigma_\epsilon^2 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}. \quad (2.24)$$

Let M_1 (M_2) represent the coefficient vector related to σ_{0k}^2 (σ_ϵ^2) in Equation 2.24, respectively. Also, let the notation vec(M) denote the stacked columns of matrix M. Then Equation 2.24 is the mean function for a formal regression model

$$\text{vec}(\tilde{Y}_k\tilde{Y}_k^T) = \text{vec}(V_k) = M_1\sigma_{0k}^2 + M_2\sigma_e^2 + R = M\theta + R \quad (2.25)$$

where $\theta = [\sigma_{0k}^2 \ \sigma_\epsilon^2]$ and R is an error term. Solving for σ_{0k}^2 and σ_ϵ^2 requires a weighted regression, where the weights are the inverse of the variances of vec($\tilde{Y}_k\tilde{Y}_k^T$).

Obtaining the M matrix in general LME Models

The details of finding the concise expression for M in a general mixed effects model are given in Section 2.6.1, leading to

$$M = \begin{bmatrix} \text{vec}(Z_1Z_1^T) & \text{vec}(Z_1Z_2^T + Z_2Z_1^T) & \dots & \text{vec}(Z_QZ_Q^T) & \text{vec}(I_N) \end{bmatrix} \quad (2.26)$$

$$M_k = \begin{bmatrix} \text{vec}(Z_{1k}Z_{1k}^T) & \text{vec}(Z_{1k}Z_{2k}^T + Z_{2k}Z_{1k}^T) & \dots & \text{vec}(Z_{Qk}Z_{Qk}^T) & \text{vec}(I_{N_k}) \end{bmatrix}, \quad (2.27)$$

where M_k is the matrix M limited to cluster k.

Unweighted solution: $\hat{\theta}$, General Mixed Effects Case

Just as a inverse-variance weighted regression was needed in Equation 2.25, it is also needed in the general case. Searle et al. (1992) §12.3bii show that $\text{var}(\text{vec}[\tilde{Y}\tilde{Y}^T]) = V \otimes V$ where $\otimes$ is the kronecker operator (note $\text{vec}(\tilde{Y}\tilde{Y}^T) = \tilde{Y} \otimes \tilde{Y}$). Thus

$$\hat{\theta} = (M^T(V^{-1} \otimes V^{-1})M)^{-1}(M^T(V^{-1} \otimes V^{-1})\text{vec}(\tilde{Y}\tilde{Y}^T)) \tag{2.28}$$

by the standard inverse variance weighted regression solution. Note that V (and V^{-1}) are block diagonal by cluster due to the structure of the Z matrix from Equation 2.1. With this, $\hat{\theta}$ can be written as functions of sums over clusters

$$\begin{aligned} \hat{\theta} &= (\sum_{k=1}^{K} M_k^T(V_k^{-1} \otimes V_k^{-1})M_k)^{-1}(\sum_{k=1}^{K} M_k^T(V_k^{-1} \otimes V_k^{-1})\text{vec}(\tilde{Y}_k\tilde{Y}_k^T)) \\ &= R^{-1}S. \end{aligned} \tag{2.29}$$

Equations 2.28 and 2.29 express $\hat{\theta}$ in terms of β through the centered variables $\tilde{Y}$. To obtain unweighted LME estimates, we provide an initial guess for $\hat{\beta}$ and $\hat{\theta}$ and iterate between estimates of β (through Equation 2.23) and θ (through Equation 2.29) until convergence. This method is equivalent to maximum likelihood estimation for LME models.

Preparing for Weights in the PSHGR Solutions

PSHGR manipulated Equations 2.23 and 2.29 into equivalent expressions where sampling weights can be inserted with the Horvitz-Thompson heuristic. Let $\mathcal{H}_{ik}$, $i = 1, ..., Q(Q+1)/2$, be matrices used to write Ω as a linear function of θ (see Equation 2.53), and let δ_{ts} be one if $\theta_t = \sigma_\epsilon^2$ and zero otherwise. Let R_{tl} be the element in the t^{th} row and l^{th} column of the matrix R (from Equation 2.29) and let S_t be the t^{th} element from matrix S (from

Equation 2.29). Section 2.6.1 below shows that

$$P = \sum_{k=1}^{K} X_k^T X_k - X_k^T \mathcal{Z}_k A_k \mathcal{Z}_k^T X_k \tag{2.30}$$

$$T = \sum_{k=1}^{K} X_k^T Y_k - X_k^T \mathcal{Z}_k A_k \mathcal{Z}_k^T Y_k \tag{2.31}$$

$$\begin{aligned} R_{tl} &= \sum_k \sigma_\epsilon^{-4} \left[\delta_{ts}\delta_{ls} N_k + \delta_{ls}\mathrm{tr}(\mathcal{Z}_k^T \mathcal{Z}_k C_{tk}) \right. \\ &+ \left. \delta_{ts}\mathrm{tr}(\mathcal{Z}_k^T \mathcal{Z}_k^T \mathcal{H}_{lk}) + \mathrm{tr}(\mathcal{Z}_k^T \mathcal{Z}_k C_{tk} \mathcal{Z}_k^T \mathcal{Z}_k \mathcal{H}_{lk})\right] \end{aligned} \tag{2.32}$$

$$S_t = \hat{\sigma}_\epsilon^{-4} \sum_k \mathrm{tr}\left(\delta_{ts} \tilde{Y}^T \tilde{Y} + \tilde{Y}^T \mathcal{Z}_k C_{tk} \mathcal{Z}_k^T \tilde{Y}\right) \tag{2.33}$$

where $t = 1, ..., s$, $l = 1, ..., s$, $s = 1, ..., Q(Q+1)/2 + 1$ (the number of unique elements in $\mathcal{O}$), and

$$\begin{aligned} C_{tk} &= -\delta_{ts} A_k + B_{tk} - B_{tk} \mathcal{Z}_k^T \mathcal{Z}_k A_k \\ B_{tk} &= \hat{\sigma}_\epsilon^2 A_k \hat{\mathcal{O}}^{-1} \mathcal{H}_{tk} - \delta_{ts} A_k \\ A_k &= (\mathcal{Z}_k^T \mathcal{Z}_k + \sigma_\epsilon^2 \mathcal{O}^{-1})^{-1}, \end{aligned}$$

and $\mathcal{Z}_k$ and $\mathcal{O}$ are defined in Equations 2.54 and 2.3. Equations 2.30 through 2.33 are ready for the insertion of the sampling weights.

Insertion of Weights into PSHGR Solutions

The Horvitz-Thompson heuristic is used to insert the cluster weights, w_k, into Equations 2.30 to 2.33. To insert the $w_{i|k}$ weights, view P, T, R, and S as functions of the census matrices $X_k^T X_k, X_k^T Y_k^T, \tilde{Y}_k^T \tilde{Y}_k, X_k^T \mathcal{Z}_k, Y_k^T \mathcal{Z}_k, \tilde{Y}_k^T \mathcal{Z}_k$ and $\mathcal{Z}_k^T \mathcal{Z}_k$, and estimate them using the Horvitz-Thompson heuristic. Let $W_{\cdot|k}$ be a diagonal matrix with elements $w_{i|k}, i = 1, ..., n_k$. Then, for example, approximate the census quantity $X_k^T \mathcal{Z}_k (\mathcal{Z}_k^T \mathcal{Z}_k^T + \sigma_\epsilon^2 \Omega^{-1})^{-1} \mathcal{Z}_k^T X_k$ from Equation 2.30 with $X_{Sk}^T W_{\cdot|k} \mathcal{Z}_{Sk} (\mathcal{Z}_{Sk}^T W_{\cdot|k} \mathcal{Z}_{Sk}^T + \hat{\sigma}_\epsilon^2 \hat{\Omega}^{-1})^{-1} \mathcal{Z}_{Sk}^T W_{\cdot|k} X_{Sk}$, which is not unbi-

ased unless independence is assumed as described in Section 2.6.1. Then

$$P_w^{(r)} = \sum_{k=1}^{K} w_k \left(X_k^T W_{\cdot|k} X_k - X_k^T W_{\cdot|k} \mathcal{Z}_k A_{kw}^{(r-1)} \mathcal{Z}_k^T W_{\cdot|k} X_k \right) \quad (2.34)$$

$$T_w^{(r)} = \sum_{k=1}^{K} w_k \left(X_k^T W_{\cdot|k} Y_k - X_k^T W_{\cdot|k} \mathcal{Z}_k A_{kw}^{(r-1)} \mathcal{Z}_k^T W_{\cdot|k} Y_k \right) \quad (2.35)$$

$$R_{tlw}^{(r)} = \sum_{k} w_k \left(\delta_{tS}\delta_{lS} N_k + \delta_{lS} \mathrm{tr}(\mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k C_{tkw}^{(r)}) + \delta_{tS} \mathrm{tr}(\mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k^T \mathcal{H}_{lk}) \right) \quad (2.36)$$

$$+ \sum_{k} w_k \left(\mathrm{tr}(\mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k C_{tkw}^{(r)} \mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k \mathcal{H}_{lk}) \right)$$

$$S_{tw}^{(r)} = \hat{\sigma}_e^{-4} \sum_{k} w_k \mathrm{tr} \left(\delta_{tS} \tilde{Y}^T W_{\cdot|k} \tilde{Y} + \tilde{Y}^T W_{\cdot|k} \mathcal{Z}_k C_{tkw}^{(r)} \mathcal{Z}_k^T W_{\cdot|k} \tilde{Y} \right) \quad (2.37)$$

where $C_{tkw}^{(r)} = -\delta_{tS} A_{kw}^{(r)} + B_{tkw}^{(r)} - B_{tkw}^{(r)} \mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k A_{kw}^{(r)}$, $B_{tkw}^{(r)} = \hat{\sigma}_e^2 A_{kw}^{(r)} \hat{\Omega}^{-1(r-1)} \mathcal{H}_{tk} - \delta_{tS} A_{kw}^{(r)}$ and $A_{kw}^{(r)} = \left(\mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k + \hat{\sigma}_\epsilon^2 \hat{\Omega}^{-1(r-1)} \right)^{-1}$. The addition of the superscript (r) indicates the r^{th} iteration. The estimates are obtained by solving for $\hat{\beta}_w^{(r)}$ and $\hat{\theta}_w^{(r)}$ at each iteration, where

$$\hat{\beta}_w^{(r)} = \left(P_w^{(r)} \right)^{-1} T_w^{(r)}$$
$$\hat{\theta}_w^{(r)} = \left(R_w^{(r)} \right)^{-1} S_w^{(r)}.$$

2.3 Methods for Weighting in LME Models – Variances of Point Estimates

In the estimation of point estimates, all of the methods start with model based maximum likelihood methods to estimate model parameters. Each method, RHS, KG and PSHGR, estimates the census likelihood quantities using weighted sample quantities, albeit in slightly different ways. For the variance of the weighted point estimates, the three methods differ in their approaches, which are expanded upon in this section.

RHS obtain a model based estimate of the standard errors from the estimated popu-

lation likelihood by using the sandwich estimator, see Binder (1983), Skinner (1989a) and Pawitan (2001). The sandwich estimator computes the variance of a maximum likelihood estimate when the likelihood is incorrectly specified. In this framework, the true model is the census likelihood, and the incorrect model the weighted sample likelihood. Rabe-Hesketh and Skrondal (2006) and Asparouhov (2006) use this likelihood based estimator, which tends to be preferred by non-survey statisticians.

KG obtain a design based estimate of the standard errors by using the jackknife method to account for the design, see Skinner (1989b) and Korn and Graubard (1999)). Survey statisticians prefer non-parametric techniques to account for the complexities due to the lack of parametric assumptions. Korn and Graubard (2003) recommend the jackknife estimate.

PSHGR formulate a full superpopulation model (incorporating both model and design-based theories) and then argue in Pfeffermann (1993) that the design-based component of variance dominates the variance in the standard scenario when N is large. Pfeffermann et al. (1998) recommend this approximation.

2.3.1 RHS Variances - Parametric Sandwich Estimators

Sandwich Estimator Derivation

Suppose that $Y_1, ..., Y_n$ are generated according to probability function $f(Y|\theta)$, but we are using $Y_1, ..., Y_n$ to estimate τ, where $g(Y|\tau)$ is a misspecified distribution. If $Y_1, ..., Y_n$ are independent, then the log likelihood of τ given g is

$$\begin{aligned} l(\tau) &= \log g(Y|\tau) \\ &= \sum_i \log g(Y_i|\tau). \end{aligned} \tag{2.38}$$

Let $\hat{\tau}$ be the value of τ maximizing this likelihood equation. With the misspecification of the model, the MLE of τ produces $g(\tau)$ as close to $f(\theta)$ as possible. To see this, consider

a new quantity, τ^*,

$$\tau^* = \underset{\tau}{\operatorname{argmax}} E_f l(\tau) \approx \underset{\tau}{\operatorname{argmax}} \frac{1}{n} \sum_i E_f \log g(Y_i|\tau) = \underset{\tau}{\operatorname{argmax}} E_f \log g(Y_1|\tau)$$

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}}\, l(\tau) = \underset{\tau}{\operatorname{argmax}} \sum_i \log g(Y_i|\tau) = \underset{\tau}{\operatorname{argmax}} \frac{1}{n} \sum_i \log g(Y_i|\tau)$$

Note that by the law of large numbers,

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}} \frac{1}{n} \sum_i \log g(Y_i|\tau) \rightarrow \underset{\tau}{\operatorname{argmax}} E_f \log g(Y_1|\tau) = \tau^* \text{ as } N \rightarrow \infty$$

where the Y_i are independently and identically distributed. Thus,

$$|\tau^* - \hat{\tau}| < \epsilon \text{ as } N \rightarrow \infty$$

The value τ^* maximizing $E_f l(\tau)$ corresponds to finding the value of τ that minimizes the Kullback-Leibler distance between the true distribution, $f(\theta)$, and the estimated distribution, $g(\tau)$. To see this, note that $E_f \log f(Y|\theta)$ is an unknown constant with respect to τ. Maximizing $E_f \log g(Y|\tau)$ then minimizes

$$D(f, g) = E_f \log f(Y|\theta) - E_f \log g(Y|\tau),$$

the Kullback-Leibler distance.

To obtain the distribution of $\hat{\tau}$, take the Taylor series expansion of $l'(\hat{\tau})$ about τ^*,

$$l'(\hat{\tau}) = 0 \;\approx\; l'(\tau^*) + (\hat{\tau} - \tau^*) l''(\tau^*),$$

where $l'(\hat{\tau}) = 0$ because $\hat{\tau}$ minimizes $l(\tau)$. To estimate the distribution of $\hat{\tau}$, we obtain an

expression for $\sqrt{n}(\hat{\tau} - \tau^*)$,

$$\sqrt{n}(\hat{\tau} - \tau^*) = \frac{\sqrt{n}^{-1} l'(\tau^*)}{-n^{-1} l''(\tau^*)}.$$

Note that, $l'(\tau^*) = \sum_i \frac{\partial}{\partial \tau} \log g(Y_i|\tau^*)$, so by the CLT, $\sqrt{n}^{-1} l'(\tau^*)$ converges in distribution to $N(E_f \frac{\partial}{\partial \tau} \log g(Y_1|\tau^*), J(\tau^*))$. Assuming regularity conditions, $E_f(l'(\tau^*)) = 0$ as τ^* maximizes $E_f l(\tau)$. For the variance, $J(\tau^*) = \text{var}_f(l'(\tau^*)) = E_f(l'(\tau^*) l'^T(\tau^*))$.

By the weak law of large numbers, $(-1)n^{-1} l''(\tau^*) = (-1)n^{-1} \sum_i \frac{\partial^2}{\partial \tau^{*2}} \log g(Y_i|\tau^*)$ converges to $I(\tau^*) = E_f(-l''(\tau^*))$. Thus,

$$\hat{\tau} - \tau* \sim_{n\to\infty} N(0, I^{-1}(\tau^*) J(\tau^*) I^{-1}(\tau^*))$$

where $I(\tau^*) = E_f(-l''(\tau^*))$ and $J(\tau^*) = E_f(l'(Y|\tau^*) l'^T(Y|\tau^*))$. Because the value of τ^* is not known, plug-in $\hat{\tau}$ for τ^* resulting in the following approximation

$$\hat{\tau} - \tau^* \sim_{n\to\infty} N(0, \hat{I}^{-1}(\hat{\tau}) \hat{J}(\hat{\tau}) \hat{I}^{-1}(\hat{\tau})) \tag{2.39}$$

where

$$\hat{I}(\hat{\tau}) = E_f(-l''(\hat{\tau})), \ \hat{J}(\hat{\tau}) = E_f(l'(Y|\hat{\tau}) l'^T(Y|\hat{\tau})). \tag{2.40}$$

If the Y_i's are not independent with respect to g, then the computation of the sandwich estimator becomes more complicated. The log likelihood in Equation 2.38 needs to be re-written to take the sampling design into account. For example, if the sampling design has H strata and within each stratum there are K clusters, $l(\tau)$ is

$$\begin{aligned} l(\tau) &= \log g(Y|\tau) \\ &= \sum_{h=1}^{H} \sum_{k=1}^{K} \log g(Y_{hk}|\tau). \end{aligned}$$

The above computations can be repeated to form a new $l'(\tau)$ and the resulting $\hat{I}(\hat{\tau})$ and $\hat{J}(\hat{\tau})$ will reflect the covariance structure of the Y_{hki}'s.

When does $I(\tau^*) = J(\tau^*)$, allowing $\hat{\tau} \sim N(0, I^{-1}(\tau^*))$? Well,

$$\begin{aligned}
l''(\tau^*) &= \frac{\partial}{\partial \tau^*}\frac{\partial}{\partial \tau^{*T}} \log g(Y|\tau^*) \\
&= \frac{\partial}{\partial \tau^*}\left[\frac{1}{g(Y|\tau^*)}\frac{\partial}{\partial \tau^*} g(Y|\tau^*)\right] \\
&= \frac{g(Y|\tau^*)\frac{\partial^2}{\partial \tau^* \partial \tau^{*T}} g(Y|\tau^*) - [\frac{\partial}{\partial \tau^*} g(Y|\tau^*)]^2}{g(Y|\tau^*)^2} \\
&= \frac{\frac{\partial^2}{\partial \tau^* \partial \tau^{*T}} g(Y|\tau^*)}{g(Y|\tau^*)} - \left[\frac{\frac{\partial}{\partial \tau^*} g(Y|\tau^*)}{g(Y|\tau^*)}\right]^2 \\
&= \frac{\frac{\partial^2}{\partial \tau^* \partial \tau^{*T}} g(Y|\tau^*)}{g(Y|\tau^*)} - \left[\frac{\partial}{\partial \tau^*} l(Y|\tau^*)\right]^2
\end{aligned}$$

which implies that

$$I(\tau^*) = E_f((-1)l''(\tau^*) \;=\; E_f[l'(\tau^*)]^2 - E_f\left(\frac{L''(\tau^*)}{L(\tau^*)}\right).$$

If $E_f\left(\frac{L''(\tau^*)}{L(\tau^*)}\right) = 0$, then $I(\tau^*) = J(\tau^*)$. Observe that,

$$\begin{aligned}
E_f\left(\frac{L''(\tau^*)}{L_g(\tau^*)}\right) &= \int \cdots \int \frac{L''(\tau^*)}{L(\tau^*)} f(Y|\theta) dy_1 \cdots dy_n \\
&= \int \cdots \int \frac{\frac{\partial^2}{\partial \tau^* \partial \tau^{*T}} g(Y|\tau^*)}{g(Y|\tau^*)} f(Y|\theta) dy_1 \cdots dy_n.
\end{aligned}$$

If $f = g$ (and $\tau^* = \theta$), then this integral becomes $\frac{\partial^2}{\partial \tau^* \tau^{*T}} 1 = 0$ and the distribution of the $\hat{\tau}$ is $N(0, I(\theta)^{-1})$, assuming regularity conditions. If $f \neq g$, the variance is the sandwich estimator variance, called so because the J term is "sandwiched" between the I terms.

RHS Implementation of the Sandwich Estimator

RHS use the sandwich estimator to estimate variances. The correctly specified distribution, $f(\theta)$, is the generating LME model incorporating the sampling design, and the incorrectly specified distribution, $g(\tau)$, is the sample weighted likelihood (joint density) function. With the LME estimation, $\tau = \{\beta, \Omega, \sigma_\epsilon^2\} = \{\beta, \theta\}$. Recall the likelihood equations in Equation 2.11 and 2.12 and the definitions of $I(\hat{\tau})$ and $J(\hat{\tau})$ from Equation 2.40. Let τ represent the pseudo-likelihood parameter vector and k_s represent the number of sampled clusters. Then

$$\begin{aligned} l'_w(\tau) &= \sum_{k=1}^{k_s} w_k \frac{\partial}{\partial \tau} \log L_w(Y_k|X, Z, \tau) \\ &\equiv \sum_{k=1}^{k_s} l'_{kw}(\tau), \end{aligned}$$

where $l'_{kw}(\tau)$ represents the weighted score function in cluster k. The unbiased estimator of J is

$$\hat{J}(\hat{\tau}) = \frac{k_s}{k_s - 1} \sum_{k=1}^{k_s} l'_{kw}(\hat{\tau}) l'^T_{kw}(\hat{\tau}).$$

The fraction $\frac{k_s}{k_s-1}$ above makes the MLE estimate unbiased. Next,

$$\hat{I}(\hat{\tau}) = -l''_w(\hat{\tau}) = \sum_{k=1}^{k_s} w_k \frac{\partial^2}{\partial \tau^2} \log L_w(Y_k|X, Z, \tau)|_{\tau=\hat{\tau}}.$$

The expected Fisher information is estimated by the observed Fisher information evaluated at the ML estimates.

The estimates of J and I are used in Equation 2.39 to obtain an estimate of the variance of the weighted point estimates.

2.3.2 KG Variances - Non-Parametric Jackknife Estimates

Korn and Graubard (2003) use jackknife techniques to compute the variance. Let $\hat{\tau}_{(k)}$ represent the estimate of $\tau = \{\beta, \Omega, \sigma_\epsilon^2\} = \{\beta, \theta\}$ when all sampled clusters except for cluster k are used to generate the estimate. Let $\hat{\tau}$ be the estimate when all sampled clusters are used to generate the estimate. Then

$$\hat{\text{var}}(\hat{\tau}) = \frac{k_s - 1}{k_s} \sum_{k=1}^{k_s} (\hat{\tau}_{(k)} - \hat{\tau})^2.$$

This can be modified to incorporate a stratified sampling design by utilizing the independence across clusters, where k_{sh} are the number of sampled clusters in stratum h,

$$\hat{\text{var}}(\hat{\tau}) = \sum_{h=1}^{H} \frac{k_{sh} - 1}{k_{sh}} \sum_{k=1}^{k_s} (\hat{\tau}_{(k)} - \hat{\tau})^2.$$

The details of jackknife estimation with survey sampling data are described in Wolter (1985), Chapter 4. He specifically addresses stratified and stratified/clustered sampling designs with inference to infinite and finite populations. A disadvantage is the intensive computational time. For more information on the jackknife procedure in survey sampling, see Sarndal et al. (1992) §11.5, and Korn and Graubard (1999).

2.3.3 PSHGR Variances - Design-Based Estimates

Pfeffermann et al. (1998) estimate the variance of the point estimates with a design-based estimate, justified in Pfeffermann (1993). Let $\hat{\tau}$ be an estimate of $\tau = \{\beta, \Omega, \sigma_\epsilon^2\} = \{\beta, \theta\}$. When computing the variance, both the randomization distribution, p, and the generating (i.e. LME) model, ξ need to be considered. Using rules of conditional variance,

$$\text{var}_{p\xi}(\hat{\tau}) = E_\xi(\text{var}_p(\hat{\tau}|Y)) + \text{var}_\xi(E_p(\hat{\tau}|Y)),$$

where $\hat{\tau}$ is the ML estimate of τ from the weighted pseudo-likelihood. Let $\hat{\tau}$, τ_c and τ be the

weighted sample estimate, the census estimate and the true parameter value, respectively. If $\hat{\tau}$ is consistent for τ_c, and τ_c is consistent for τ, then $\hat{\tau}$ is consistent for τ. For a mean, the convergence of $\hat{\tau}$ to τ_c is $O_p(n^{-\frac{1}{2}})$ and convergence of τ_c to τ is $O_p(N^{-\frac{1}{2}})$. So

$$\begin{aligned}\hat{\tau}-\tau &= (\hat{\tau}-\tau_c)+(\tau_c-\tau)\\ &= O_p(n^{-\frac{1}{2}})+O_p(N^{-\frac{1}{2}})\\ &= O_p(n^{-\frac{1}{2}}).\end{aligned}$$

We can re-write the variance decomposition as

$$\begin{aligned}\mathrm{var}_{p\xi}(\hat{\tau}) &= E_\xi(\mathrm{var}_p(\hat{\tau}|Y))+\mathrm{var}_\xi(E_p(\hat{\tau}|Y))\\ &= E_\xi(\mathrm{var}_p(\hat{\tau}|Y))+O_p(N^{-1}).\end{aligned}$$

The variance is then estimated as $\mathrm{var}_p(\hat{\tau}|Y)$ as N^{-1} is quite small for most populations.

PSHGR Implementation

The PSHGR design-based variance estimate of $\hat{\beta}$ uses Equations 2.34 and 2.35 in a multivariate Taylor series approximation. Following, for example, Sarndal et al. (1992) §5.12,

$$\begin{aligned}\hat{\beta} &\approx \beta+\sum_{j=1}^{p}\sum_{k<j}\frac{\partial\hat{\beta}}{\partial\hat{P}_{jk}}|_{\hat{P}=P,\hat{T}=T}(\hat{P}_{jk}-P_{jk})+\sum_{j=1}^{p}\frac{\partial\hat{\beta}}{\partial\hat{T}_j}|_{\hat{P}=P,\hat{T}=T}(\hat{T}_{jo}-T_{jo})\\ &= \beta+P^{-1}(\hat{T}-\hat{P}\beta).\end{aligned} \tag{2.41}$$

The details are in Section 2.6.2. The random elements in this equation are the hatted matrices, so our initial expression for $\mathrm{Var}(\hat{\beta})$ is

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}(\beta+P^{-1}(\hat{T}-\hat{P}\beta))\\ &= P^{-1}\mathrm{Var}(\hat{T}-\hat{P}\beta)P^{-1}.\end{aligned}$$

Substituting in the estimates, $\hat{T}$ and $\hat{P}$ from Equations 2.34 and 2.35, we get that

$$\text{Var}(\hat{\beta}) = P^{-1}\text{Var}(\sum_k w_k c_k)P^{-1}$$

where

$$c_k = \left(X_k^T W_{\cdot|k} Y_k - X_k^T W_{\cdot|k} Z_k A_{kw} Z_k^T W_{\cdot|k} Y_k\right) - \left(X_k^T W_{\cdot|k} X_k - X_k^T W_{\cdot|k} Z_k A_{kw} Z_k^T W_{\cdot|k} X_k\right)\beta.$$

Assume the PSU inclusion probabilities are iid (approximately true when the sampling fraction is small). Then

$$\begin{aligned}
\text{var}(\hat{\beta}) &= P^{-1}\text{Var}(\sum_{k=1}^{k_s} w_k c_k)P^{-1} \\
&= P^{-1}\sum_{k=1}^{k_s}\text{Var}(w_k c_k)P^{-1}, \text{ independent} \\
&= P^{-1}k_s\text{Var}(w_k c_k)P^{-1}, \text{ for any } k, \text{ identically distributed.} \qquad (2.42)
\end{aligned}$$

Note that $\sum_k w_k c_k = P\beta - T$ where P and T are defined in Equations 2.34 and 2.35. Observe that $P\beta - T = P(P^{-1}T) - T = 0$ so that

$$\text{Var}(w_k c_k) = \frac{1}{k_s - 1}\sum_k w_k^2 c_k c_k^T. \qquad (2.43)$$

Substituting Equation 2.43 into Equation 2.42, we obtain the PSGHR estimate of the variance of β,

$$\begin{aligned}
\text{Var}(\hat{\beta}) &= P^{-1}k_s\text{Var}(w_k c_k)P^{-1}, \text{ for any } k \\
&= P^{-1}\frac{k_s}{k_s - 1}\left(\sum_k w_k^2 c_k c_k^T\right)P^{-1}. \qquad (2.44)
\end{aligned}$$

The estimates for the variance of $\hat{\theta}$ is obtained in a similar manner. Using the values

of R and S from Equations 2.36 and 2.37

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= \mathrm{Var}\left(\theta + R^{-1}(\hat{S} - \hat{R}\theta)\right) \\
&= R^{-1} k_s \mathrm{Var}(w_k d_k) R^{-1}, \text{ for any } k \\
&= R^{-1} \frac{k_s}{k_s - 1} \sum_k w_k^2 d_k d_k^T R^{-1} \\
&= R^{-1} \frac{k_s}{k_s - 1} \left(\sum_k w_k^2 d_k d_k^T \right) R^{-1} \qquad (2.45)
\end{aligned}$$

where

$$\begin{aligned}
d_k &= \mathrm{tr}\left(\delta_{tS} \tilde{Y}^T W_{\cdot|k} \tilde{Y} + \tilde{Y}^T W_{\cdot|k} \mathcal{Z}_k C_{tkw} \mathcal{Z}_k^T W_{\cdot|k} \tilde{Y}\right) - (\delta_{tS}\delta_{lS} N_k \\
&+ \delta_{lS} \mathrm{tr}(\mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k C_{tkw}) + \delta_{tS} \mathrm{tr}(\mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k^T \mathcal{H}_{lk}) - \mathrm{tr}(\mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k C_{tkw} \mathcal{Z}_k^T W_{\cdot|k} \mathcal{H}_{lk})) \theta \\
C_{tkw} &= -\delta_{tS} A_{kw} + B_{tkw} - B_{tkw} \mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k A_{kw} \\
B_{tkw} &= \hat{\sigma}_e^2 A_{kw} \hat{\Omega}^{-1} \mathcal{H}_{tk} - \delta_{tS} A_{kw} \\
A_{kw} &= \left(\mathcal{Z}_k^T W_{\cdot|k} \mathcal{Z}_k + \hat{\sigma}_\epsilon^2 \hat{\Omega}^{-1}\right)^{-1}.
\end{aligned}$$

Next, we estimate $\mathrm{Var}(\hat{\beta})$ and $\mathrm{Var}(\hat{\theta})$ from Equations 2.44 and 2.45 by substituting in the estimates of P and R.

2.4 Consistency, Scaling and Comparisons

2.4.1 Consistency of PML Estimates for LME Models

The definitions and conditions for asymptotic evaluation under both the randomization distribution and the ξ distribution are described in Section 1.2.4. This section shows that for the PML β estimates derived in this chapter to be consistent the number of clusters sampled needs to increase as the sample size increases. For the estimates of the elements in the variance matrix, σ_ϵ^2 and ω from Equation 2.3, both the number of clusters and the

number of elements per cluster need to increase. This consistency argument is shown in detail below for RHS. The arguments for KG and PSHGR follow similarly and are not shown.

As a consequence of the consistency argument below, it is possible to multiply the conditional weights, $w_{i|k}$'s, by a constant dependent on the cluster to reduce the bias in the variance components. These computations are shown in Section 2.4.2. A direct comparison of the PSHGR and RHS methods for a random intercept model shows conditions necessary for the two methods to equal each other in Section 2.4.3.

RHS Consistency

Recall from Section 1.2.4 and Sarndal et al. (1992) the scenario in which consistency is derived in survey samples. Let N be an infinite sequence of elements, and let $U_1, U_2, ...$ represent a sequence of populations where U_m consists of the first N_m elements from the sequence N. Let $N_1 < N_2 < \cdots$. so that $U_1 \subset U_2 \subset \cdots$. Let θ_m be a population parameter of interest so that θ_p is function of $y_1, ..., y_{N_m}$. For each population, U_p, consider the sampling design, p_m that assigns probability $p_m(s_m)$ to each sample s_m from U_m. Assume that n_m is the sample size for population U_m and that $m_1 < m_2 < \cdots$. Clearly, $m \to \infty$ implies $N_m \to \infty$ and $n_m \to \infty$. The way in which the sample size and the population grow depends upon the specific estimator and sampling design being evaluated. Below the consistency for RHS is shown for the weighted LME estimators under a clustered design.

Using the RHS method, the number of clusters needs to increase for consistency of the β estimates, and the number of clusters and the number of elements per cluster needs to increase for the consistency of the θ estimates. To see this,we begin by deriving the PML

estimates $\hat{\beta}$ and $\hat{\theta}$ from the RHS weighted likelihood from Equation 2.13,

$$\begin{aligned} L_w(Y|X,\beta,V^{-1}) &= \prod_{k=1}^{k_s} L_w(Y_k|X,\beta,V^{-1})^{w_k} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^{\hat{N}} \left(\prod_{k=1}^{k_s} |V_{kw}^{-1}|^{w_k}\right) \\ &\times \exp\left(-\frac{1}{2}\sum_{k=1}^{k_s} w_k \left\{(Y_k - X_k\beta)^T V_{kw}^{-1}(Y_k - X_k\beta)\right\}\right). \end{aligned}$$

We find the derivative of the log likelihood with respect to β

$$\frac{\partial}{\partial\beta} l_w(Y_k|X,\beta,V^{-1}) = \sum_{k=1}^{k_s} w_k \left(X_k^T V_{kw}^{-1} Y_k\right) - \sum_{k=1}^{k_s} w_k \left(X_k^T V_{kw}^{-1} X_k\beta\right)$$

and estimate $\hat{\beta}$ by setting the score equal to zero,

$$\hat{\beta}_w = \left[\sum_{k=1}^{k_s} w_k \left(X_k^T V_{kw}^{-1} X_k\right)\right]^{-1} \sum_{k=1}^{k_s} w_k \left(X_k^T V_{kw}^{-1} Y_k\right)$$

where

$$V_{kw}^{-1} = \sigma_\epsilon^{-2}\left[W_{\cdot|k} - W_{\cdot|k}\mathcal{Z}_k A_k \mathcal{Z}_k^T W_{\cdot|k}\right]. \tag{2.46}$$

We will show consistency by showing that the bias and variance of $\hat{\beta}_w$ go to zero as the number of sampled clusters increases. Assume that the elements of V_{kw}^{-1} are known. Let

$$\begin{aligned} T_1 &= \sum_{k=1}^{K} \left(X_{kC}^T V_{kC}^{-1} X_{kC}\right), \quad \hat{T}_1 = \sum_{k=1}^{k_s} w_k \left(X_k^T V_{kw}^{-1} X_k\right) \\ T_2 &= \sum_{k=1}^{K} \left(X_{kC}^T V_{kC}^{-1} Y_{kC}\right), \quad \hat{T}_2 = \sum_{k=1}^{k_s} w_k \left(X_k^T V_{kw}^{-1} Y_k\right) \end{aligned}$$

As the number of sampled clusters increases to the population cluster size, $\hat{T}_1 \to T_1$ and $\hat{T}_2 \to T_2$. Taylor series expansion of $\hat{\beta}_w$ about the census estimate of β, β_C, see Sarndal

et al. (1992) §5.12, provides the following approximation

$$\hat{\beta}_w \approx \beta_C + T_1^{-1}(\hat{T}_2 - \hat{T}_1\beta_C) \quad (2.47)$$

$$- \left[T_1^{-1}(\hat{T}_1 - T_1)T_1^{-1}(\hat{T}_2 - T_2) + 2T_1^{-1}(\hat{T}_1 - T)T_1^{-1}(\hat{T}_1^{-1} - T_1)\beta_C\right]. \quad (2.48)$$

For the terms of the approximation of $\hat{\beta}_w$ in Equation 2.47, we see that $E(\hat{\beta}_w) = \beta_C + T_1^{-1}(T_2 - T_1\beta_C) = \beta_C$ since $T_1^{-1}T_2 = \beta_C$. For the terms in Equation 2.48, we need to assume that T_1^{-1} gets smaller faster than $(\hat{T}_1 - T_1)$ and $(\hat{T}_2 - T_2)$ get larger. We know that T_1 is strictly increasing because V_{kw}^{-1} is positive definite. With these assumptions, $\hat{\beta}_w$ is asymptotically unbiased for β_C. We assume that the finite population parameter β_C is unbiased (or asymptotically unbiased) for β. From this, $\hat{\beta}_w$ is asymptotically unbiased for β.

We next look at the variance. Recall $\mathrm{Var}_{p\xi}(\hat{\beta}_w) = E_\xi(\mathrm{Var}_p(\hat{\beta}_w|Y)) + \mathrm{Var}_\xi(E_p(\hat{\beta}_w|Y))$ and

$$\begin{aligned} \mathrm{Var}_\xi(E_p(\hat{\beta}_w|Y)) &= \mathrm{Var}_\xi(\beta_c) = T_1^{-1} \\ \mathrm{Var}_p(\hat{\beta}_w|Y) &\approx T_1^{-1}\mathrm{Var}_p(\hat{T}_2 - \hat{T}_1\beta_c)T_1^{-1} \end{aligned}$$

We can obtain an expression for $\mathrm{Var}(\hat{T}_2 - \hat{T}_1\beta_c)$ by assuming that the number of clusters sampled is small compared to the number of population clusters. This makes the inclusion variables approximately independent, giving us

$$\mathrm{Var}_p(\hat{\beta}_w|Y) \approx T_1^{-1}\sum_{k=1}^{K}\left(w_k(X_k^TV_{kw}^{-1}[Y_k - X_k\beta_c])\right)\mathrm{Var}_p(I_k)\left(w_k(X_k^TV_{kw}^{-1}[Y_k - X_k\beta_c])\right)^T T_1^{-1}$$

This sum is written over the number of population clusters, however the inclusion indicator (I_k) reduces the sum to be over the number of sampled clusters. Recall that $\mathrm{Var}(I_k) =$

$\pi_k - \pi_k^2 = \frac{1}{w_k} - \frac{1}{w_k^2}$. Next take the expected value with respect to ξ to obtain

$$E_\xi \mathrm{Var}_p(\hat{\beta}_w|Y) \approx T_1^{-1}\left(\sum_{k=1}^{K}(w_k-1)X_k^T V_{kw}^{-1}X_k\right)T_1^{-1},$$

under the assumption that the I_k's are independent of one another. Pulling these together, we get an expression for the variance,

$$\mathrm{Var}_{p\xi}(\hat{\beta}_w) \approx T_1^{-1} + T_1^{-1}\left(\sum_{k=1}^{K}(w_k-1)X_k^T V_{kw}^{-1}X_k\right)T_1^{-1}$$

For consistency, we need $T_1^{-1} \overset{k\to\infty}{\to} 0$ and $T_1^{-1}\left(\sum_{k=1}^{K}(w_k-1)X_k^T V_{kw}^{-1}X_k\right)T_1^{-1} \overset{k\to\infty}{\to} 0$. This is reasonable because V_{kw}^{-1} is positive definite.

To obtain the variance with respect to elements in V_{kw}, a secondary regression to estimate the elements of V_{kw} in terms of a weighted regression estimator, as described in Section 2.4.1, provides

$$\hat{\theta}_w = \left(\sum_k w_k M_{kw}^T(V_{kw}^{-1}\otimes V_{kw}^{-1})M_{kw}\right)^{-1}\left(\sum_k w_k M_{kw}^T(V_{kw}^{-1}\otimes V_{kw}^{-1})\mathrm{vec}(\hat{\hat{Y}}_{kw}\hat{\hat{Y}}_{kw}^T)\right).$$

Next assume that the β values are known (recall $\hat{\hat{Y}}_{kw} = Y_k - X_k\hat{\beta}_w$). Unlike the consistency argument for $\hat{\beta}_w$, the effect of changing the sample size on V_{kw}^{-1} elements needs to be considered as the elements of V_{kw}^{-1} are not known. Similar to the previous argument, let

$$T_1^* = \sum_k M_{kC}^T(V_{kC}^{-1}\otimes V_{kC}^{-1})M_{kC}^T, \qquad \hat{T}_1^* = \sum_k w_k M_{kw}^T(V_{kw}^{-1}\otimes V_{kw}^{-1})M_{kw}$$

$$T_2^* = \sum_k M_{kC}^T(V_{kC}^{-1}\otimes V_{kC}^{-1})\mathrm{vec}(Y_{kC},Y_{kC}^T), \qquad \hat{T}_2^* = \sum_k w_k M_{kw}^T(V_{kw}^{-1}\otimes V_{kw}^{-1})\mathrm{vec}(\hat{\hat{Y}}_{kw}\hat{\hat{Y}}_{kw}^T)$$

A Taylor series expansion similar to Equation 2.47 holds for $\hat{\theta}_w$,

$$\hat{\theta}_w \approx \theta + T_1^{*-1}(\hat{T}_2^* - \hat{T}_1^*\theta_c).$$

we need $\hat{T}_1^* \to T_1^*$ and $\hat{T}_2^* \to T_2^*$ for $\hat{\theta}_w$ to be asymptotically unbiased. For this to happen, we need the number of clusters to increase as we did for $\hat{\beta}_w$. However, we also need to make sure that there is no systematic bias in V_{kw}^{-1}. Recall that

$$V_{kw}^{-1} = \sigma_\epsilon^{-2}\left[W_{\cdot|k} - W_{\cdot|k}\mathcal{Z}_k A_k \mathcal{Z}_k^T W_{\cdot|k}\right]$$

when

$$\begin{aligned} A_k &= (\mathcal{Z}_k^T W_{\cdot|k}\mathcal{Z}_k + \sigma_\epsilon^2\Omega^{-1})^{-1} \\ &= (\sum_i w_{i|k}\mathcal{Z}_{1k}^T\mathcal{Z}_k + \sigma_\epsilon^2\Omega^{-1})^{-1}. \end{aligned}$$

V_{wk} is a function of A_k, which may have some systematic bias in the estimation of $\sum_{i=1}^{n_k} w_{i|k}\mathcal{Z}_k^T\mathcal{Z}$. The only way to remove this systematic bias is to increase the number of elements per cluster. For the variance components, both the number of clusters and the number of elements per cluster need to increase as the sample size increases.

If the sampling design has strata along with clusters, then the structure of the X and $\mathcal{Z}$ matrices will change. To ensure consistency in these cases, the conditions stated here need to hold. Specifically, $T_1^{-1} \overset{k\to\infty}{\to} 0$, $T_1^{-1}\left(\sum_{k=1}^K (w_k - 1)X_k^T V_{kw}^{-1} X_k\right) T_1^{-1} \overset{k\to\infty}{\to} 0$ and V_{kw}^{-1} must have no systematic bias.

2.4.2 Scaling of the Weights

The PML methods in this chapter require that both the number of clusters and the number of elements per cluster increase for consistency of the variance estimates. However, the consistency arguments in the previous subsection are invariant to scaling the conditional weights (i.e. $w_{i|k}$) by a constant dependent on the cluster. We can take advantage of this to reduce the bias in the variance components. Scaling of the cluster weights is not considered as it has no effect on the estimates, see Equations 2.13 for RHS, Equations 2.17 to 2.19 for KG and Equations 2.34 through 2.37 for PSHGR.

To demonstrate the effect of scaling the weights on estimation bias, we adapt and extend an example from Searle et al. (1992) §3.7 using a *balanced* random intercept model with no covariates. To adapt this to the survey setting, assume that the estimated number of population elements in each cluster is the same, or $\sum_{i|k} w_{i|k} = \hat{N}_1$ (the size of the first cluster) for all k. The model is

$$y_{ik} = \beta_0 + U_{0k} + \epsilon_{ik}, \quad U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon). \tag{2.49}$$

From Searle et al. (1992) we know that the maximum likelihood estimates for the unweighted σ^2_{0k} and σ^2_ϵ have a closed form. Let k_s be the number of sampled clusters and n be the number of individuals sampled in each cluster. Then the MLE's are

$$\begin{aligned} \hat{\sigma}^2_\epsilon &= \sum_k \sum_i \frac{(y_{ik} - \bar{y}_{\cdot k})^2}{k_s(n-1)} \\ \hat{\sigma}^2_a &= \frac{\sum_k (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2}{k_s} - \frac{\hat{\sigma}^2_\epsilon}{n}. \end{aligned}$$

When these estimators are derived using psuedo-maximum likelihood (such as the RHS or PSHGR methods), we get

$$\begin{aligned} \hat{\sigma}^2_{w\epsilon} &= \sum_k w_k \sum_i w_{i|k} \frac{(y_{ik} - \bar{y}_{\cdot k})^2}{\hat{K}(\hat{N}_1 - 1)} \\ \hat{\sigma}^2_{w0k} &= \frac{\sum_k w_k (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2}{\hat{K}} - \frac{\hat{\sigma}^2_{w\epsilon}}{\hat{N}_1}, \end{aligned}$$

where $\hat{K} = \sum_k w_k$ is the number of estimated population clusters. The expected values of

these estimators (with respect to the model in Equation 2.49) is

$$E_\xi(\hat{\sigma}^2_{w\epsilon}) = \sigma^2_\epsilon \left(1 + \frac{\sum_k w_k (1 - \frac{\sum_i w^2_{i|k}}{\hat{N}_1})}{\hat{K}(\hat{N}_1 - 1)} \right) \tag{2.50}$$

$$E_\xi(\hat{\sigma}^2_{w0k}) = \sigma^2_{0k} \left(1 - \frac{\sum_k w^2_k}{\hat{K}^2} \right) \tag{2.51}$$

$$+ \frac{\sigma^2_\epsilon}{\hat{N}_1} \left(\frac{\sum_k w_k \sum_i w^2_{i|k}}{\hat{K}\hat{N}_1} - \frac{\sum_k w^2_k \sum_i w^2_{i|k}}{\hat{K}^2 \hat{N}_1} - \frac{\sum_k w_k (1 - \frac{\sum_i w^2_{i|k}}{\hat{N}_1})}{\hat{K}(\hat{N}_1 - 1)} - 1 \right) \tag{2.52}$$

Notice that when the analysis is unweighted ($w_{i|j} = 1 \; \forall i, j$), then the estimate of σ^2_ϵ is unbiased. However, the unweighted estimate of σ^2_{0k} has a bias of $\frac{1}{\hat{K}}\sigma^2_{0k} - \frac{1}{\hat{K}\hat{N}_1}\sigma^2_\epsilon$.

PSHGR and RHS use two scaling methods. For the first scaling, called *Weighted Scaled 1* by PSHGR , when $\frac{\sum_i w^2_{i|k}}{\hat{N}_1} = 1$ in Equations 2.50 to 2.52, $\hat{\sigma}^2_{w\epsilon}$ is unbiased and $\hat{\sigma}^2_{w0k}$ has a bias of $\frac{\sum_k w^2_k}{\hat{K}^2}\sigma^2_{0k} - \frac{\sum_k w^2_k}{\hat{N}_1\hat{K}^2}\sigma^2_\epsilon$. Using $\frac{\sum_i w^2_{i|k}}{\hat{N}_1} = 1$ implies a form for the conditional weights under the scaled 1 method, $w^{s1}_{i|j} = w_{i|j}\lambda_k$. Setting $\frac{\sum_i (w^{s1}_{i|k})^2}{\sum_i w^{s1}_{i|k}} = 1$ is equivalent to $\frac{\sum_i w^2_{i|k}\lambda^2_k}{\sum_i w_{i|k}\lambda_k} = 1$, implying $\lambda_k = \frac{\sum_i w_{i|k}}{\sum_i w^2_{i|k}}$. Thus,

$$w^{(s1)}_{i|k} = w_{i|k} \frac{\sum_i w_{i|k}}{\sum_i w^2_{i|k}}.$$

Note $\sum_i w^{(s1)}_{i|k} = \sum_i w_{i|k} \frac{\sum_j w_{j|k}}{\sum_j w^2_{j|k}} = \frac{\left(\sum_i w_{i|k}\right)^2}{\sum_i w^2_{i|k}}$, which Potthoff et al. (1992) call the equivalent sample size. Potthoff et al. (1992) note that $\sum_i w^{(s1)}_{i|k} = \sum_i \left(w^{(s1)}_{i|k} \right)^2$ and demonstrate that the equivalent sample size acts as a sample size in a Horvitz-Thompson style mean estimation example.

For the second scaling, called *Weighted Scaled 2* by PSHGR, the estimated population size is scaled to be the sample size. Let n_k be the sample size for cluster k. Then

$$w^{(s2)}_{i|k} = w_{i|k} \frac{n_k}{\sum_i w_{i|k}}$$

Clearly, $\sum_i w_{i|k}^{(s2)} = n_k$.

To compare these scalings, consider again the random intercept model and the biases defined by Equation 2.49 through 2.52. Table 2.1 summarizes these results.

- *Weighted Unscaled*: Using $\sum_i w_{i|k} = \hat{N}_1$, $\sum_k w_k = \hat{K}$ and $\hat{N}_1 \leq \sum_i w_{i|k}^2 \leq \hat{N}_1^2$ in Equation 2.50, the bias for $\hat{\sigma}_{w\epsilon}^2$ can be bounded, $-\sigma_\epsilon^2 \leq E(\hat{\sigma}_{w\epsilon}^2) - \sigma_\epsilon^2 \leq 0$. The lower bound will be achieved when all individuals per cluster are sampled so conditional weights equal one, and the upper bound will be reached when only one individual per cluster is sampled.

  Similarly, using $K \leq \sum_k w_k^2 \leq K^2$, the bias for σ_{0k}^2 can be bounded also. By looking at the four conditions where $\sum_i w_{i|k}^2$ is $\hat{N}_1$ or $\hat{N}_1^2$ and $\sum_k w_k^2$ is $\hat{K}$ or $\hat{K}^2$, bounds on the bias are obtained and listed in Table 2.1. Given the weights, the bias of σ_{w0k}^2 is dependent on the intra-class correlation ($\frac{\sigma_{0k}^2}{\sigma_{0k}^2+\sigma_\epsilon^2}$), If $\sigma_\epsilon^2 << \sigma_{0k}^2$ then the bias will be negative. However, if $\sigma_{0k}^2 << \sigma_\epsilon^2$, then the bias may become positive if $\sum_k w_k^2 = K$ and $\sum_i w_{i|k}^2 = N_1^2$. Though it was shown in a limited case here, this bias is also exhibited in many of the simulations in Section 3.4 whose designs are not balanced.

- *Weighted Scaled 1*: For the scaled 1 weights, we have that $w_{i|k}^{s1} = w_{i|k}\frac{\sum_i w_{i|k}}{\sum_i w_{i|k}^2} = w_{i|k}\frac{\hat{N}_1}{\sum_i w_{i|k}^2}$. From this, we see that $1 \leq \sum_i w_{i|k}^{s2} \leq \hat{N}_1$ and $\sum_i (w_{i|k}^{s2})^2 = \sum_i w_{i|k}^{s2}$. The bounds are displayed in Table 2.1.

  This is related to the ICC in the same manner as the weighted unscaled case.

- *Weighted Scaled 2*: For the scaled 2 weights, we have that $w_{i|k}^{s2} = w_{i|k}\frac{n_1}{\sum_i w_{i|k}} = w_{i|k}\frac{n_1}{\hat{N}_1}$. Thus $\frac{n_1^2}{\hat{N}_1} \leq \sum_i (w_{i|k}^{s2})^2 \leq n_1^2$. The bounds are displayed in Table 2.1. The relationship between the bias and the ICC depends on the sampling fraction.

From Table 2.1, the bias for $\hat{\sigma}_{w\epsilon}^2$ is always negative for the unscaled weights, and always positive for the scaled 1 weights. The range of bias for the the scaled 1 weights is smaller than for the unscaled weights. As $n_1 \to \hat{N}_1$ then the bounds with the scaled 2 weights

Method	Bound on $\hat{\sigma}^2_{w\epsilon}$ Bias
Unscaled	$-\sigma^2_\epsilon \leq \text{bias} \leq 0$
Scaled 1	$0 \leq \text{bias} \leq \frac{\sigma^2_\epsilon}{\hat{N}_1}$
Scaled 2	$\frac{\sigma^2_\epsilon(\hat{N}_1-n_1^2)}{\hat{N}_1(\hat{N}_1-1)} \leq \text{bias} \leq \frac{\sigma^2_\epsilon(\hat{N}_1^2-n_1^2)}{\hat{N}_1^2(\hat{N}_1-1)}$

Table 2.1: Bias for $\hat{\sigma}^2_{w\epsilon}$ Under Different Weighting Methods

Method	Value for $(\sum_k w_k)(\sum_i w_{i\|k})$	$\hat{\sigma}^2_{wok}$ Bias
Unscaled	$(K)(N_1)$	$-\frac{1}{KN_1}\sigma^2_\epsilon - \frac{1}{K}\sigma^2_{0k}$
	$(K)(N_1^2)$	$\left(1-\frac{1}{K}\right)\sigma^2_\epsilon - \frac{1}{K}\sigma^2_{0k}$
	$(K^2)(N_1)$	$-\frac{1}{N_1}\sigma^2_\epsilon - \sigma^2_{0k}$
	$(K^2)(N_1^2)$	$-\sigma^2_{0k}$
Scaled 1	$(K)(N_1)$	$-\frac{1}{KN_1}\sigma^2_\epsilon - \frac{1}{K}\sigma^2_{0k}$
	$(K)(N_1^2)$	$-\left(\frac{1}{N_1}+\frac{1}{KN_1^2}\right)\sigma^2_\epsilon - \frac{1}{K}\sigma^2_{0k}$
	$(K^2)(N_1)$	$-\frac{1}{N_1}\sigma^2_\epsilon - \sigma^2_{0k}$
	$(K^2)(N_1^2)$	$-\left(\frac{1}{N_1}+\frac{1}{N_1^2}\right)\sigma^2_\epsilon - \sigma^2_{0k}$
Scaled 2	$(K)(N_1)$	$\left(\frac{n_1^2}{N_1^3}-\frac{n_1^2}{KN_1^3}-\frac{N_1^2-n_1^2}{N_1^3(N_1-1)}-\frac{1}{N_1}\right)\sigma^2_\epsilon - \frac{1}{K}\sigma^2_{0k}$
	$(K)(N_1^2)$	$\left(\frac{n_1^2}{N_1^2}-\frac{n_1^2}{KN_1^2}-\frac{N-n_1^2}{N_1^2(N_1-1)}-\frac{1}{N_1}\right)\sigma^2_\epsilon - \frac{1}{K}\sigma^2_{0k}$
	$(K^2)(N_1)$	$-\left(\frac{N_1^2-n_1^2}{N_1^3(N_1-1)}+\frac{1}{N_1}\right)\sigma^2_\epsilon - \sigma^2_{0k}$
	$(K^2)(N_1^2)$	$-\left(\frac{N_1-n_1^2}{N_1^2(N_1-1)}+\frac{1}{N_1}\right)\sigma^2_\epsilon - \sigma^2_{0k}$

Table 2.2: Bias for $\hat{\sigma}^2_{w0k}$ Under Different Weighting Methods

approach $-\sigma_\epsilon^2 \leq$ bias ≤ 0 and as $n_1 \rightarrow 1$ then the bounds approach $\frac{\sigma_\epsilon^2}{N_1} \leq$ bias $\leq \frac{\sigma_\epsilon^2}{N_1}$. If $n_1 = 1$, σ_ϵ^2 can not be estimated, hence the artificially small range on the bias.

From Table 2.2, the bounds on the bias depend on the relationship between σ_ϵ^2 and σ_{0k}^2. For example, in the weighted unscaled case, if $\sigma_{0k}^2 >> \sigma_\epsilon^2$ then the lower bound will be $-\sigma_{0k}^2$ and the upper bound will be $-\frac{1}{K}\sigma_{0k}^2$. If $\sigma_{0k}^2 << \sigma_\epsilon^2$, then the lower bound is $-\frac{1}{N_1}\sigma_\epsilon^2$ and the upper bound is $(1-\frac{1}{K})\sigma_\epsilon^2$. In general, if $\sigma_{0k}^2 >> \sigma_\epsilon^2$, then all of the methods have a lower bound on the bias of $-\sigma_{0k}^2$ and an upper bound of $-\frac{1}{K}\sigma_{0k}^2$. If $\sigma_\epsilon^2 >> \sigma_{0k}^2$ then the bounds get more complex. The lower bounds for the unscaled, scaled1 and scaled 2 estimates are all on the order of $-\frac{1}{N_1}\sigma_\epsilon^2$. The upper bounds are $(1-\frac{1}{K})\sigma_\epsilon^2, -\frac{1}{KN_1}\sigma_\epsilon^2$ and $-\frac{1}{N_1}\sigma_\epsilon^2$ respectively. From this, the unscaled weighting is the only one that can have a positive upper bound on the bias. The range of the bias for the scaled 1 weightings is smaller than for the scaled 2 weightings.

While this analysis was done for the case of a random intercept only model, the trends will be seen in the more complex models simulated in the next chapter. Specifically, for $\hat{\sigma}_{w\epsilon}^2$ the unscaled bias is larger and negative, the scaled 1 bias is smaller and positive and the scaled 2 bias tends to be in between them (varies according to sampling proportion) and for $\hat{\sigma}_{w0k}^2$ the lower bound of the bias is similar for all scalings, and the upper bound is largest for the unscaled weights and smallest for the scaled 1 weights.

2.4.3 Comparisons between Methods

This section compares the PSHGR approach and the RHS approach for a random intercept model. Including the KG approach in this comparison is difficult due to the different weights used. This section comments on the uniqueness of the KG model and then compares the RHS and PSHGR methods.

KG

KG requires a different set of weights than the other approaches from this chapter. Recall that w_k is the inverse probability that cluster k is included in the sample, and that $w_{i|k}, w_{ij|k}, w_{ijs|k}w_{ijst|k}$ are the inverse probabilities that elements i, ij, ijs and $ijst$, respectively, in cluster k are included in the sample given that cluster k is in the sample. It is unlikely that secondary analysts will have access to the higher order conditional weights, making this approach not practical. Even researchers who have access to the extended (non-public) use datasets will rarely have this information.

RHS vs. PSHGR

The RHS and PSHGR approaches appear quite different from their derivations. In this section, the weighted RHS estimates are compared to the weighted PSHGR estimates for a random intercept model. We show below that for a random intercept model where the estimated number of population elements for each cluster is the same, then the two methods are similar. When estimated number of population elements for each cluster varies, then the RHS and PSHGR estimates differ more substantially.

The random intercept model is $Y_{ik} = X_{ik}\beta + U_{0k} + \epsilon_{ik}$ where $U_{0k} \sim N(0, \sigma^2_{0k}), \epsilon_{ik} \sim N(0, \sigma^2_\epsilon)$ letting $\lambda_k = N_k\sigma^2_{0k} + \sigma^2_\epsilon$. To estimate the parameters, compute the RHS weighted likelihood for each cluster,

$$\begin{aligned} L(Y_k|X, \beta, V^{-1}) &= \int \left[\prod_i L(y_{ik}|U_{0k}, \sigma^2_\epsilon, x_{ik})^{w_{i|k}}\right] L(U_{0k}|\sigma^2_{0k})dU_{0k} \\ &= \left[\frac{1}{\sqrt{2\pi}\sqrt{\sigma^2_\epsilon}}\right]^{N_k} \frac{\sqrt{\sigma^2_\epsilon}}{\sqrt{\lambda_k}} \\ &\times \exp\left(\frac{-1}{2\sigma^2_\epsilon}\sum_i w_{i|k}(y_{ik} - x_{ik}\beta)^2 + \frac{\sigma^2_{0k}}{2\sigma^2_\epsilon\lambda_k}\left[\sum_i w_{i|k}(y_{ik} - x_{ik}\beta)\right]^2\right) \end{aligned}$$

The full likelihood can be computed as

$$\begin{aligned} L(Y|X,\beta,V^{-1}) &= \prod_k L(Y_k|X,\beta,V^{-1})^{w_k} \\ &= \left(\frac{1}{\sqrt{2\pi}\sqrt{\sigma_\epsilon^2}}\right)^N \frac{(\sigma_\epsilon^2)^{\frac{K}{2}}}{\prod_k \lambda_k^{\frac{w_k}{2}}} \\ &\times \exp\left(-\sum_k \frac{w_k}{2\sigma_\epsilon^2}\left[\sum_i w_{i|k}(y_{ik}-x_{ik}\beta)^2 - \frac{\sigma_{0k}^2}{\lambda_k}\left(\sum_i w_{i|k}(y_{ik}-x_{ik}\beta)\right)^2\right]\right) \end{aligned}$$

Finally, the log likelihood can be computed,

$$\begin{aligned} l(Y|X,\beta,V^{-1}) &= -\frac{N}{2}\log(2\pi) - \frac{N-K}{2}\log(\sigma_\epsilon^2) - \sum_k \frac{w_k}{2}\log(\lambda_k) \\ &- \frac{\sum_k w_k}{2\sigma_\epsilon^2}\left[\sum_i w_{i|k}(y_{ik}-x_{ik}\beta)^2 - \frac{\sigma_{0k}^2}{\lambda_k}\left(\sum_i w_{i|k}(y_{ik}-x_{ik}\beta)\right)^2\right]. \end{aligned}$$

For computational ease, substitute $\frac{\lambda_k-\sigma_\epsilon^2}{N_k}$ for σ_{0k}^2 in the log likelihood,

$$\begin{aligned} l(Y|X,\beta,V^{-1}) &= -\frac{N}{2}\log(2\pi) - \frac{N-K}{2}\log(\sigma_\epsilon^2) - \sum_k \frac{w_k}{2}\log(\lambda_k) \\ &- \frac{\sum_k w_k}{2\sigma_\epsilon^2}\left[\sum_i w_{i|k}(y_{ik}-x_{ik}\beta)^2 - \frac{\lambda_k-\sigma_\epsilon^2}{N_k\lambda_k}\left(\sum_i w_{i|k}(y_{ik}-x_{ik}\beta)\right)^2\right]. \end{aligned}$$

Let $\hat{N}_k = \sum_i w_{i|k}$, $\hat{K} = \sum_k w_k$, $\bar{y}_{\cdot kw} = \frac{1}{\hat{N}_k}\sum_i w_{i|k}y_{ik}$, and $\bar{x}_{\cdot kw} = \frac{1}{\hat{N}_k}\sum_i w_{i|k}x_{ik}$. The

parameter estimates are

$$\begin{aligned}
\hat{\beta}_{\text{RHS}} &= \frac{\sum_k w_k \sum_i w_{i|k} y_{ik}(x_{ik} - \frac{\hat{\lambda}_k - \sigma^2_\epsilon}{\hat{\lambda}_k}\bar{x}_{\cdot kw})}{\sum_k w_k \sum_i w_{i|k} x_{ik}(x_{ik} - \frac{\hat{\lambda}_k - \sigma^2_\epsilon}{\hat{\lambda}_k}\bar{x}_{\cdot kw})} \\
\hat{\sigma}^2_{\epsilon}\text{RHS} &= \frac{1}{\hat{N} - \hat{K}} \sum_k w_k \left[\sum_k w_{i|k}(y_{ik} - x_{ik}\beta)^2 - \hat{N}_k(\bar{y}_{\cdot kw} - \bar{x}_{\cdot kw}\beta)^2 \right] \\
\hat{\lambda}_k\text{RHS} &= \hat{N}_k(\bar{y}_{\cdot kw} - \bar{x}_{\cdot kw}\beta)^2 \\
\hat{\sigma}^2_{0k}\text{RHS} &= \sum_k w_k \frac{\hat{\lambda}_k - \hat{\sigma}^2_\epsilon}{\hat{K}\hat{N}_k}.
\end{aligned}$$

It is easily shown that if $\hat{N}_k = \hat{N}_1$ (which implies $\hat{\lambda}_k = \hat{\lambda}$), $x_{ik} = 1$ and all weights equal 1, these reduce to the estimates for the unweighted random intercept model with no covariates given in Searle et al. (1992).

Now we return to the PSHGR estimates. Allowing $\mathcal{Z}_k$ to be a vector of ones, use Equations 2.34 to 2.37 to compute

$$\begin{aligned}
\hat{P}_w &= \sum_k w_k \left[\sum_i w_{i|k} x_{ik} x_{ik}^T - \frac{\hat{\sigma}^2_{0k}}{\hat{\lambda}_k} (\sum_i w_{i|k} x_{ik})(\sum_i w_{i|k} x_{ik})^T \right] \\
\hat{T}_w &= \sum_k w_k \left[\sum_i w_{i|k} x_{ik} y_{ik} - \frac{\hat{\sigma}^2_{0k}}{\hat{\lambda}_k} (\sum_i w_{i|k} x_{ik})(\sum_i w_{i|k} y_{ik})^T \right] \\
\hat{R}_w &= \begin{bmatrix} \sum_k w_k \frac{\hat{N}_k^2}{\hat{\lambda}_k^2} & \sum_k w_k \frac{\hat{N}_k}{\hat{\lambda}_k^2} \\ \sum_k w_k \frac{\hat{N}_k}{\hat{\lambda}_k^2} & \sum_k w_k \left[\frac{\hat{N}_k - 1}{\hat{\sigma}^4_\epsilon} + \frac{1}{\lambda_k^2} \right] \end{bmatrix} \\
\hat{S}_w &= \begin{bmatrix} \sum_k w_k \frac{\hat{N}_k^2}{\hat{\lambda}_k^2} \left[\bar{y}_{\cdot kw} - \bar{x}_{\cdot kw}\hat{\beta} \right]^2 \\ \sum_k w_k \left[\sum_i \frac{w_{i|k}}{\hat{\sigma}^4_\epsilon}(y_{ik} - x_{ik}\hat{\beta})^2 - \left(\frac{\hat{N}_k}{\hat{\sigma}^4_\epsilon} - \frac{N_k}{\hat{\lambda}_k^2} \right) (\bar{y}_{\cdot kw} - \bar{x}_{\cdot kw}\hat{\beta})^2 \right] \end{bmatrix}
\end{aligned}$$

Substituting $\frac{\hat{\lambda}_k - \sigma^2_\epsilon}{\hat{N}_k}$ for σ^2_{0k} in $\hat{P}$ and $\hat{T}$ above, we see that

$$\hat{\beta}_{\text{PSHGR}} = \left(\hat{P}_w \right)^{-1} \hat{T}_w = \hat{\beta}_{\text{RHS}}.$$

To get the variance components, note that $[\hat{\sigma}^2_{0k} \;\; \hat{\sigma}^2_\epsilon] = \hat{\theta} = \left(\hat{R}_w \right)^{-1} \hat{S}_w$ provide the following

estimates

$$\begin{aligned}
\hat{\sigma}^2_{0k} &= \frac{1}{C}\left[\left(\sum_k w_k\left[\frac{(\hat{N}_k-1)}{\hat{\sigma}^4_\epsilon}+\frac{1}{\hat{\lambda}^2_k}\right]\right)\left(\sum_k w_k\frac{\hat{N}^2_k}{\hat{\lambda}^2_k}\left[\bar{y}_{\cdot kw}-\bar{x}_{\cdot kw}\hat{\beta}\right]^2\right)\right] \\
&- \frac{1}{C}\left[\left(\sum_k w_k\frac{\hat{N}_k}{\hat{\lambda}^2_k}\right)\left(\sum_k \frac{w_k}{\hat{\sigma}^4_\epsilon}\left[\sum_i w_{i|k}(y_{ik}-x_{ik}\hat{\beta})^2-\left(\hat{N}_k-\frac{\hat{\sigma}^4_\epsilon\hat{N}_k}{\hat{\lambda}^2_k}\right)(\bar{y}_{\cdot kw}-\bar{x}_{\cdot kw}\hat{\beta})^2\right]\right)\right] \\
C &= \left(\sum_k w_k\frac{\hat{N}^2_k}{\hat{\lambda}^2_k}\right)\left(\sum_k w_k\left[\frac{\hat{N}_k-1}{\hat{\sigma}^4_\epsilon}+\frac{1}{\hat{\lambda}^2_k}\right]\right)-\left(\sum_k w_k\frac{\hat{N}_k}{\hat{\lambda}^2_k}\right)^2 \\
\hat{\sigma}^2_\epsilon &= \frac{1}{C}\left(-\sum_k w_k\frac{\hat{N}_k}{\hat{\lambda}^2_k}\right)\left(\sum_k w_k\frac{\hat{N}^2_k}{\hat{\lambda}^2_k}\left[\bar{y}_{\cdot kw}-\bar{x}_{\cdot kw}\hat{\beta}\right]^2\right) \\
&+ \frac{1}{C}\left(\sum_k w_k\frac{\hat{N}^2_k}{\hat{\lambda}^2_k}\right)\left(\sum_k w_k\left[\sum_i \frac{w_{i|k}}{\hat{\sigma}^4_\epsilon}(y_{ik}-x_{ik}\hat{\beta})^2-\left(\frac{\hat{N}_k}{\hat{\sigma}^4_\epsilon}-\frac{\hat{N}_k}{\hat{\lambda}^2_k}\right)(\bar{y}_{\cdot kw}-\bar{x}_{\cdot kw}\hat{\beta})^2\right]\right).
\end{aligned}$$

The PSHGR estimates can be re-written as a function of the estimate from RHS,

$$\begin{aligned}
\hat{\sigma}^2_{\epsilon}\text{PSHGR} &= \frac{1}{\hat{C}}\left(\sum_k w_k\frac{\hat{N}^2_k}{\hat{\lambda}^2_k}\right)\left(\sum_k w_k\frac{1}{\hat{\lambda}_k}\right)-\frac{1}{\hat{C}}\left(\sum_k w_k\frac{\hat{N}_k}{\hat{\lambda}^2_k}\right)\left(\sum_k w_k\frac{\hat{N}_k}{\hat{\lambda}_k}\right) \\
&+ \frac{1}{\hat{\sigma}^4_\epsilon\hat{C}}\left(\sum_k w_k\frac{\hat{N}^2_k}{\hat{\lambda}^2_k}\right)\left(\hat{\sigma}^2_{\epsilon RHS}\right)\left(\hat{N}-\hat{K}\right)
\end{aligned}$$

where

$$\begin{aligned}
\hat{C} &= \left(\sum_k w_k\frac{\hat{N}^2_k}{\hat{\lambda}^2_k}\right)\frac{\left(\hat{N}-\hat{K}\right)}{\hat{\sigma}^4_\epsilon}+\left(\sum_k w_k\frac{\hat{N}^2_k}{\hat{\lambda}^2_k}\right)\left(\sum_k\frac{w_k}{\hat{\lambda}^2_k}\right)-\left(\sum_k w_k\frac{\hat{N}_k}{\hat{\lambda}^2_k}\right)^2 \\
\hat{\sigma}^2_{0k}\text{PSHGR} &= \frac{\hat{N}-\hat{K}}{\hat{\sigma}^4_\epsilon\hat{C}}\left(\sum_k w_k\frac{\hat{N}_k}{\hat{\lambda}^2_k}\left(\hat{\lambda}_k-\hat{\sigma}^2_\epsilon\text{RHS}\right)\right) \\
&- \frac{1}{\hat{C}}\left(\sum_k w_k\frac{\hat{N}_k}{\hat{\lambda}^2_k}\right)\left(\sum_k w_k\frac{1}{\hat{\lambda}_k}\right)+\frac{1}{\hat{C}}\left(\sum_k\frac{w_k}{\hat{\lambda}^2_k}\right)\left(\sum_k w_k\frac{\hat{N}_k}{\hat{\lambda}_k}\right)
\end{aligned}$$

First consider the case where $\hat{N}_k = \hat{N}_1$ for all k. Then $\hat{\lambda}_k = \hat{N}_1\hat{\sigma}^2_{0k} + \hat{\sigma}^2_\epsilon = \hat{\lambda}$ and the first two terms of $\hat{\sigma}^2_\epsilon\text{PSHGR}$ cancel out, as do the second and third term of $\hat{C}$. With this, $\hat{\sigma}^2_\epsilon\text{PSHGR} = \hat{\sigma}^2_\epsilon\text{RHS}$. For $\hat{\sigma}^2_{0k}\text{PSHGR}$, when $\hat{N}_k = \hat{N}_1$ then the last two terms cancel each other out. The first term (with the reduction in $\hat{C}$ mentioned above) becomes $\hat{N}_1^{-1}\hat{\lambda}_\text{RHS}$, and the second term becomes $\frac{1}{\hat{N}_1}\hat{\sigma}^2_\epsilon\text{RHS}$. From this, $\hat{\sigma}^2_{0k}\text{PSHGR} = \hat{\sigma}^2_{0k}\text{RHS}$. Thus we have that in the balanced case, the PSHGR and RHS estimates in a random intercept model

are identical.

2.5 Summary

RHS, KG and PSHGR insert sample weights into LME estimates by using the PML method at different stages of the estimation, resulting in three different sets of estimators. All the estimators are consistent for the β coefficients if the number of sampled clusters increases as the sample size increases, however both the number of sampled clusters and the number of sampled elements per cluster need to increase for the θ, the random effect variances, to be consistent. By scaling the conditional weights (the $w_{i|k}$) the bias in the random effect variances can be reduced. The KG method is different from RHS and PSHGR as KG require higher order conditional weights that are not usually available to secondary analysts. Finally, it was shown that the RHS and PSHGR methods are the same for a random intercept model where the estimated population size for each cluster is the same, and different for random intercept models when the estimated number of population elements per cluster are different.

2.6 Appendix to Chapter 2

2.6.1 PSHGR Weighting Details from Section 2.2.5

Solving for the M matrix in LME Models

The form for the matrix M is shown for a random intercept model in Equations 2.24 and 2.25 . We now compute M for a general LME model from Equation 2.2. Let $\tilde{Y} = Y - X\beta$ and $\tilde{Y}_k = Y_k - X_k\beta$. We need to express the Ω matrix from Equation 2.3 as a linear function of the individual elements of ω in Ω. Recall the Ω matrix is of dimension $KQ \times KQ$, but with only $Q(Q+1)/2$ unique w elements. Let θ be a vector of $S = Q(Q+1)/2+1$ elements containing the unique elements in Ω and the random error variance σ^2_ϵ. Then

$$
\begin{aligned}
\Omega &= \omega_{11}\begin{bmatrix} I_{KxK} & 0_{KxK} & \cdots & 0_{KxK} \\ 0_{KxK} & 0_{KxK} & \cdots & 0_{KxK} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{KxK} & 0_{KxK} & \cdots & 0_{KxK} \end{bmatrix} + \omega_{12}\begin{bmatrix} 0_{KxK} & I_{KxK} & \cdots & 0_{KxK} \\ I_{KxK} & 0_{KxK} & \cdots & 0_{KxK} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{KxK} & 0_{KxK} & \cdots & 0_{KxK} \end{bmatrix} \\
&+ \cdots + \omega_{QQ}\begin{bmatrix} 0_{KxK} & 0_{KxK} & \cdots & 0_{KxK} \\ 0_{KxK} & 0_{KxK} & \cdots & 0_{KxK} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{KxK} & 0_{KxK} & \cdots & I_{KxK} \end{bmatrix} \\
&= \sum_{i=1}^{Q}\sum_{j\geq i}^{Q} \omega_{ij} H_{ij} \\
&= \sum_{s=1}^{S-1} \theta_s H_s,
\end{aligned}
$$

where H_{ij} is a $KQ \times KQ$ matrix where the ij^{th} and ji^{th} $K \times K$ block is the identity matrix, and the other $K \times K$ blocks are zero. Let H_s represent the H_{ij} block for the

specific element in θ for which $\theta_s = \omega_{ij}$. From Equation 2.4, Var(Y) is

$$\begin{aligned} V &= Z\Omega Z^T + \sigma_\epsilon^2 I_N \\ &= Z\left(\sum_{s=1}^{S-1} H_s\theta_s\right) Z^T + \theta_S I_{N\times N} \\ &= \sum_{s=1}^{S-1} ZH_sZ^T\theta_s + \theta_S I_{N\times N}. \end{aligned}$$

Next we compute vec(V),

$$\text{vec}(V) = \text{vec}\left(Z\left(\sum_{s=1}^{S-1} H_s\theta_s\right) Z^T\right) + \theta_S I_{N\times N}.$$

Let $H = [\text{vec}(H_1) \;\; \text{vec}(H_2) \;\; \cdots \;\; \text{vec}(H_{S-1}) \;\; \text{vec}(I_{N\times N})]$. Using the identity vec$(ABC^T) = (C \otimes A)vec(B)$ (Searle (1982) Section 12.9), where $\otimes$ is the kronecker product,

$$\begin{aligned} \text{vec}(V) &= (Z \otimes Z)\sum_{s=1}^{S-1} \text{vec}(H_s)\theta_s + \theta_S \text{vec}(I_{N\times N}) \\ &= \Big[\; (Z \otimes Z)H \;\Big]\theta \\ &= M\theta. \end{aligned}$$

So M can be written as

$$\begin{aligned} M &= (Z \otimes Z)\, H \\ &= \Big[\; \text{vec}(ZH_1Z) \;\; \text{vec}(ZH_2Z) \;\; \ldots \;\; \text{vec}(ZH_{S-1}Z) \;\; \text{vec}(ZI_{N\times N}Z) \;\Big] \\ &= \Big[\; \text{vec}(Z_1Z_1^T) \;\; \text{vec}(Z_1Z_2^T + Z_2Z_1^T) \;\; \ldots \;\; \text{vec}(Z_QZ_Q^T) \;\; \text{vec}(I_N) \;\Big], \end{aligned}$$

which matches Equation 2.26. An expression for M_k can be obtained by starting with vec(V_k) and proceeding similarly, resulting in Equation 2.27. This is described in depth in the next section.

Solving for θ in an Unweighted LME Model

Let $\mathcal{O}$ represent the variance/covariance matrix for the random effects in cluster k. Analogous to Equation 2.24, we can write the variance matrix as a sum of individual elements in the matrix

$$\begin{aligned}\mathcal{O} &= \omega_{11}\begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix} + \omega_{12}\begin{bmatrix}0 & 1 & \cdots & 0\\ 1 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix} + \cdots + \omega_{QQ}\begin{bmatrix}0 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{bmatrix}\\ &= \sum_{i=1}^{Q}\sum_{j\geq i}^{Q}\omega_{ij}\mathcal{H}_{ij}\\ &= \sum_{s=1}^{S-1}\theta_s\mathcal{H}_s, \end{aligned} \tag{2.53}$$

where $\mathcal{H}_{ij}$ is a $Q \times Q$ matrix where the ij^{th} and ji^{th} element is one, and the other elements are zero. Let $\mathcal{H}_s$ represent the $\mathcal{H}_{ij}$ matrix for the specific element in θ for which $\theta_s = \omega_{ij}$. Next, define $\mathcal{Z}_k$ as

$$\mathcal{Z}_k = \begin{bmatrix}Z_{11k} & Z_{12k} & \cdots & Z_{1Qk}\\ Z_{21k} & Z_{22k} & \cdots & Z_{2Qk}\\ \vdots & \vdots & \ddots & \vdots\\ Z_{N_k1k} & Z_{N_k2k} & \cdots & Z_{N_kQk}\end{bmatrix} = \begin{bmatrix} Z_{1k} & Z_{2k} & \cdots & Z_{Qk}\end{bmatrix}, \tag{2.54}$$

where Z_{iqk} and Z_{qk} are the same as those in Equation 2.1. Define V_k as

$$\begin{aligned} V_k &= \mathcal{Z}_k \Omega_k \mathcal{Z}_k^T + \sigma_\epsilon^2 I_{N_k \times N_k} \\ &= \sum_{s=1}^{S-1} \mathcal{Z}_k \mathcal{H}_s \mathcal{Z}_k^T \theta_s + \theta_S I_{N_k \times N_k} \\ \text{vec}(V_k) &= (\mathcal{Z}_k \otimes \mathcal{Z}_k) \sum_{s=1}^{S-1} \text{vec}(\mathcal{H}_s)\theta_s + \theta_S \text{vec}(I_{N_k \times N_k}) \\ &= M_k \theta \end{aligned}$$

where M_k is defined as in Equation 2.27

$$M_k = \left[\begin{array}{cccc} \text{vec}(Z_{1k}Z_{1k}^T) & \text{vec}(Z_{1k}Z_{2k}^T + Z_{2k}Z_{1k}^T) & \ldots & \text{vec}(Z_{Qk}Z_{Qk}^T) \quad \text{vec}(I_{N_k \times N_k}) \end{array} \right].$$

A secondary regression analogous to that in Equation 2.25 is performed, resulting in an estimate for θ,

$$\begin{aligned} \hat{\theta} &= (M^T(V^{-1} \otimes V^{-1})M)^{-1}(M^T(V^{-1} \otimes V^{-1})\text{vec}(\tilde{Y}\tilde{Y}^T)) \\ &= (\sum_{k=1}^{K} M_k^T(V_k^{-1} \otimes V_k^{-1})M_k)^{-1}(\sum_{k=1}^{K} M_k^T(V_k^{-1} \otimes V_k^{-1})\text{vec}(\tilde{Y}_k\tilde{Y}_k^T)) \\ &= (R)^{-1}S, \end{aligned}$$

which is also Equation 2.29. Equations 2.23 and 2.29 are iterated until convergence to obtain unweighted estimates of the LME parameters.

Preparing for Weights in the PSHGR Solutions

To use the Horvitz-Thompson heuristic to insert weights in the PSHGR solutions, Equations 2.30 to 2.33 need to be written as functions of linear statistics. Searle (1982) (§12.9 Theorem 3) show that $(\text{vec}(Z))^T(C \otimes C^T)\text{vec}(Z) = \text{tr}(CZ^TCZ)$ for appropriately sized matrices Z and C. PSHGR use this to get an equivalent expression for the matrix R. Note that the tl^{th} element of $S \times S$ matrix R is $\sum_{k=1}^{K} M_{tk}^T(V_k^{-1} \otimes V_k^{-1})M_{lk}$, where $\theta_t = \omega_{ij}$ and $M_{tk} = \text{vec}[(Z_{ik}Z_{jk}^T + \delta_{i\neq j}Z_{jk}Z_{ik}^T)\delta_{s\neq S} + \delta_{s=S}I_{N_k}]$. Let $G_{sk} = \text{vec}^{-1}(M_{sk})$. Using the theorem from Searle, we get that the tl^{th} element of R is $\sum_k \text{tr}(V_k^{-1}G_{tk}V_k^{-1}G_{lk})$.

Similar V^{-1} and A in Equations 2.5 and 2.6, V_k^{-1} and A_k can be defined as

$$V_k^{-1} = \sigma_\epsilon^{-2}(I_{N_k\times N_k} - \mathcal{Z}_k A_k \mathcal{Z}_k^T) \tag{2.55}$$

where

$$A_k = (\mathcal{Z}_k^T\mathcal{Z}_k + \sigma_\epsilon^2\mathcal{O}^{-1})^{-1} \tag{2.56}$$

Using these,

$$\begin{aligned} V_k^{-1}G_{tk} &= \sigma_\epsilon^{-2}(I - \mathcal{Z}_k A_k \mathcal{Z}_k)(\mathcal{Z}_k\mathcal{H}_{tk}\mathcal{Z}_k + I_{N_k}\delta_{tS}) \\ &= \sigma_\epsilon^{-2}\delta_{t=S}I_{N_k} + \sigma_\epsilon^{-2}\mathcal{Z}_k B_{tk}\mathcal{Z}_k^T \end{aligned} \tag{2.57}$$

where $B_{tk} = \hat{\sigma}_\epsilon^2 A_k \hat{\Omega}_t^{-1} \mathcal{H}_{tk} - \delta_{t=S} A_k$. Next we compute

$$\begin{aligned} \mathrm{tr}(V_k^{-1} G_{tk} V_k^{-1} G_{lk}) &= \hat{\sigma}_\epsilon^{-4} \mathrm{tr}\left(\left(\delta_{t=S} I_{N_k} + \mathcal{Z}_k B_{tk} \mathcal{Z}_k^T \right) \left(\delta_{l=S} I_{N_k} + \mathcal{Z}_k B_{lk} \mathcal{Z}_k^T \right) \right) \\ &= \hat{\sigma}_\epsilon^{-4} \mathrm{tr}\left(\left(\delta_{tS} I_{N_k} + \mathcal{Z}_k C_{tk} \mathcal{Z}_k^T \right) \left(\delta_{lS} I_{N_k} + \mathcal{Z}_k \mathcal{H}_{lk} \mathcal{Z}_k^T \right) \right) \end{aligned}$$

where $C_{tk} = -\delta_{t=S} A_k + B_{tk} - B_{tk} \mathcal{Z}_k^T \mathcal{Z}_k A_k$. Multiplying this through and using the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, provides

$$\begin{aligned} R_{tl}^{(r)} &= \sum_k \sigma_\epsilon^{-4} \left[\delta_{tS} \delta_{lS} N_k + \delta_{lS} \mathrm{tr}(\mathcal{Z}_k^T \mathcal{Z}_k C_{tk}) \right. \\ &+ \left. \delta_{tS} \mathrm{tr}(\mathcal{Z}_k^T \mathcal{Z}_k^T \mathcal{H}_{lk}) + \mathrm{tr}(\mathcal{Z}_k^T \mathcal{Z}_k C_{tk} \mathcal{Z}_k^T \mathcal{Z}_k \mathcal{H}_{lk}) \right], \end{aligned}$$

which is Equation 17 in Pfeffermann et al. (1998). Similarly the t^{th} element of S is

$$\begin{aligned} S_t &= \sum_{k=1}^{K} M_{tk}^T (V_k^{-1} \otimes V_k^{-1}) \mathrm{vec}(\tilde{Y}_k \tilde{Y}_k^T) \\ &= \sum_k \mathrm{tr}(V_{jr}^{-1} G_{tk} V_{kr}^{-1} \tilde{Y} \tilde{Y}^T). \end{aligned}$$

Using Equation 2.57, we get that

$$V_k^{-1} G_{tk} V_k^{-1} \tilde{Y} \tilde{Y}^T = \hat{\sigma}_\epsilon^{-4} \left(\delta_{t=S} I_{N_k} + \mathcal{Z}_k B_{tk} \mathcal{Z}_k^T \right) \left(I_{N_k} - \mathcal{Z}_k A_k \mathcal{Z}_k^T \right) \tilde{Y} \tilde{Y}^T.$$

Thus,

$$S_t^{(r)} = \hat{\sigma}_\epsilon^{-4} \sum_k \mathrm{tr}\left(\delta_{tS} \tilde{Y}^T \tilde{Y} + \tilde{Y}^T \mathcal{Z}_k C_{tk} \mathcal{Z}_k^T \tilde{Y} \right),$$

which is the equation after Equation 17 in Pfeffermann et al. (1998).

Insertion of Weights into PSHGR Solutions

Consider P, T, R, and S from Equations 2.30 to 2.33 as functions of the census quantities $X_k^T X_k, X_k^T Z_k, Y_k^T Z_k$, and $Z_k^T Z_k$. Substitute V_k^{-1} from Equation 2.55 into P and T,

$$\begin{aligned} P &= \sum_{k=1}^{K} X_k^T X_k - X_k^T Z_k A_k Z_k^T X_k \\ T &= \sum_{k=1}^{K} X_k^T X_k - X_k^T Z_k A_k Z_k^T Y_k \end{aligned}$$

Insert sampling weights into the census quantities via the Horvitz-Thompson heuristic,

$$\begin{aligned} P_w &= \sum_{k=1}^{K} w_k \left(X_k^T W_{\cdot|k} X_k - X_k^T W_{\cdot|k} Z_k A_{kw} Z_k^T W_{\cdot|k} X_k \right) \\ T_w &= \sum_{k=1}^{K} w_k \left(X_k^T W_{\cdot|k} Y_k - X_k^T W_{\cdot|k} Z_k A_{kw} Z_k^T W_{\cdot|k} Y_k \right). \end{aligned}$$

where $W_{\cdot|k} = \text{diag}(w_{1|k}, w_{2|k}, ..., w_{n_k|k})$, and $A_{kw} = \left(Z_k^T W_{\cdot|k} Z_k + \hat{\sigma}_e^2 \Omega^{-1} \right)^{-1}$, which is the weighted estimate of A_k from Equation 2.56. Similarly for R and S we obtain

$$\begin{aligned} R_{tlw} &= \sum_k w_k \left(\delta_{tS}\delta_{lS} N_k + \delta_{lS} \text{tr}(Z_k^T W_{\cdot|k} Z_k C_{tkw}) + \delta_{tS} \text{tr}(Z_k^T W_{\cdot|k} Z_k^T \mathcal{H}_{lk}) \right) \\ &+ \sum_k w_k \left(\text{tr}(Z_k^T W_{\cdot|k} Z_k C_{tkw} Z_k^T W_{\cdot|k} Z_k \mathcal{H}_{lk}) \right) \\ S_{tw} &= \hat{\sigma}_e^{-4} \sum_k w_k \text{tr} \left(\delta_{tS} \tilde{Y}^T W_{\cdot|k} \tilde{Y} + \tilde{Y}^T W_{\cdot|k} Z_k C_{tkw} Z_k^T W_{\cdot|k} \tilde{Y} \right) \end{aligned}$$

where $C_{tkw} = -\delta_{tS}A_{kw} + B_{tkw} - B_{tkw}\mathcal{Z}_k^T W_{\cdot|k}\mathcal{Z}_k A_{kw}$, $B_{tkw} = \hat{\sigma}_e^2 A_{kw}\hat{\Omega}^{-1}\mathcal{H}_{tk} - \delta_{tS}A_{kw}$. The estimates are obtained by solving for $\beta_w^{(r)}$ and $\theta_w^{(r)}$ at each iteration, where

$$\begin{aligned}\hat{\beta}_w^{(r)} &= \left(P_w^{(r)}\right)^{-1} T_w^{(r)} \\ \hat{\theta}_w^{(r)} &= \left(R_w^{(r)}\right)^{-1} S_w^{(r)}.\end{aligned}$$

2.6.2 PSHGR Variance Details from Section 2.3.3

Linearization of $\hat{\beta}$ for variance calculation

Recall from Equation 2.41 that the Taylor series estimate for $\hat{\beta}$ is

$$\begin{aligned}\hat{\beta} &\approx \beta + \sum_{j=1}^{p}\sum_{k<j} a_{jk}(\hat{P}_{jk} - P_{jk}) + \sum_{j=1}^{p} a_{j0}(\hat{T}_{jo} - T_{jo}) \\ &= \beta + P^{-1}(\hat{T} - \hat{P}\beta).\end{aligned}$$

To derive this, let Ω_{jk} be a matrix of zeros, with one's in the jk and kj locations, and λ_j be a vector of zeros, with a one at the j entry. Then,

$$\begin{aligned}a_{jk} &= \frac{\partial\hat{\beta}}{\partial\hat{P}_{jk}}|_{\hat{P}=P,\hat{T}=T} \\ &= (-\hat{P}^{-1}\frac{\partial P}{\partial P_{jk}}\hat{P}^{-1})\hat{T}|_{\hat{P}=P,\hat{T}=T} \\ &= -P^{-1}\Omega_{jk}\beta \\ a_{jo} &= \frac{\partial\hat{\beta}}{\partial\hat{T}_j}|_{\hat{P}=P,\hat{T}=T} \\ &= \hat{P}^{-1}\lambda_j|_{\hat{P}=P,\hat{T}=T} \\ &= P^{-1}\lambda_j\end{aligned}$$

Pulling this together, we find an estimate of $\hat{\beta}$.

$$\begin{aligned}\hat{\beta} &\approx \beta + \sum_{j=1}^{p}\sum_{k<j} -P^{-1}\Omega_{jk}\beta(\hat{P}_{jk} - P_{jk}) + \sum_{j=1}^{q} P^{-1}\lambda_j(\hat{T}_{j0} - T_{jo}) \\ &= \beta - P^{-1}(\hat{P} - P)\beta + P^{-1}(\hat{T} - T) \\ &= \beta + P^{-1}(\hat{T} - \hat{P}\beta)\end{aligned}$$

2.6.3 RHS Consistency Details from Section 2.4.1

RHS Secondary Regression

When V_{kw}^{-1} is defined as in Equation 2.46, then $V_{kw}V_{kw}^{-1} = I$ when

$$V_{kw} = \left[\mathcal{Z}_k \Omega \mathcal{Z}_k^T W_{\cdot|k} + \sigma_\epsilon^2\right] W_{\cdot|k}^{-1}$$

Performing a secondary regression as in Equations 2.24 and 2.25, we obtain

$$\begin{aligned}\text{vec}(V_{wk}) &= \text{vec}(\mathcal{Z}_k \Omega \mathcal{Z}_k^T) + \sigma_\epsilon^2 \text{vec}(W_{\cdot|k}^{-1}) \\ &= \text{vec}(\sum_s \mathcal{Z}_k H_s \mathcal{Z}_k^T \theta) + \sigma_\epsilon^2 \text{vec}(W_{\cdot|k}^{-1})\end{aligned}$$

This allows us to define M_{kw} as

$$M_{kw} = \left[\begin{array}{cccccc} \text{vec}(Z_{1k}Z_{1k}^T) & \text{vec}(Z_{1k}Z_{2k}^T + Z_{2k}Z_{2k}^T) & \cdots & \text{vec}(Z_{Qk}Z_{Qk}^T) & \text{vec}(W_{\cdot|k}^{-1}) \end{array}\right]$$

From this, we get an estimate for the elements of the variance matrix,

$$\hat{\theta}_w = \left(\sum_k w_k M_{kw}^T (V_{kw}^{-1} \otimes V_{kw}^{-1}) M_{kw}\right)^{-1} \left(\sum_k w_k M_{kw}^T (V_{kw}^{-1} \otimes V_{kw}^{-1}) \text{vec}(\tilde{Y}_{kw} \tilde{Y}_{kw}^T)\right)$$

Chapter 3

Sampling Weights, Model Misspecification and Informative Sampling: A Simulation Study

This chapter evaluates the claims from Chapters 1 and 2, and evaluates them on simulated data using the methods from Chapter 2. The specific goals are: 1) To compare the results from the different methods of inserting weights into LME models, 2) To compare the sandwich estimator of the variance of the point estimates to the design-based estimator, 3) To compare the results that use different scalings of the weights, 4) To investigate the assertion that adding sampling weights can compensate for informative sampling in LME models and 5) To investigate the assertion that adding sampling weights can compensate for model misspecification in LME models. Results of the simulation studies are presented for side-by-side comparisons of parameter estimates under different simulated conditions.

Section 3.2 summarizes the previous simulation studies, including their designs and results. Section 3.3 provides a description of the format of the new simulation results presented in this dissertation. Section 3.4 describes and presents results from the 12 new simulations. Section 3.5 compares the simulations with respect to a mean squared error metric. Section 3.6 summarizes the results from the 12 new simulations and explain how these new results verify and expand the previous simulation results. Finally, Section 3.7 contains a technical appendix.

The main contribution of this chapter are the 12 simulation sets and the conclusions from them. There are four main conclusions based on these simulations. 1) The PSHGR and RHS point estimates are very similar. The differences in the point estimates are due to numerical instabilities in the estimation procedures. 2) Confidence intervals based on the sandwich estimator and the design based estimator of the variances provide similar coverage when there is no model misspecification. However, when there is model misspecification, the design-based variance estimator has unexpectedly large coverage, implying that the variance estimates are too large. 3) When there is model misspecification that does not induce informative sampling, weighted estimates do not reduce bias of the estimators. 4) When there is informative sampling, the weighted estimators do reduce the bias of the point estimates, though they do not eliminate it. 5) The unweighted estimate has the smallest variance. When there is informative sampling, the unweighted estimates are biased. The weighted unscaled estimate corrects the bias in the fixed effects, but produces more bias in the random effects. The scaled 1 weightings remove the bias in the fixed effects, and overcorrect for the weighted unscaled bias in the random effects. The scaled 2 weightings

remove the bias in the fixed effects and are in between the weighted unscaled and weighted scaled 1 bias in the random effects.

3.1 Simulation Goals and Summary of Results

As mentioned above, there are five specific goals for this chapter. In this section I describe each of them and provide a summary of the results from the simulations.

The first goal is to compare the results from the different methods of inserting weights into LME models. Recall from Chapter 2 that there are three published methods on inserting weights into LME models, Rabe-Hesketh and Skrondal (2006), denoted RHS, Korn and Graubard (2003), denoted KG, and Pfeffermann et al. (1998), denoted PSHGR. These methods use pseudo-maximum likelihood methods and differ in the location during the maximum likelihood estimation where the census quantities are estimated with weighted sample quantities (see Section 2.2). As noted in Section 2.2.3, Asparouhov (2006), denoted ASP, published the same procedure as RHS at the same time. I focus on the RHS method, as opposed to the ASP method, as the software to implement RHS was available to me whereas the software to implement ASP was not available to me. The simulations in this chapter compare the RHS and PSGHR methods, as the KG method requires additional weights that are generally not available, see Section 2.2.4. These simulations found that the RHS and PSHGR methods provide remarkably similar results. The differentiation between the methods is that the software that implements RHS (the `gllamm()` function in Stata) is not always numerically stable. This is due to the numerical quadrature implemented for the RHS method. For more details on the numerical instabilities, see Section 3.7.3.

The second goal is to compare the sandwich estimator (used by RHS, see Section 2.3.1) and the design-based estimator (used by PSHGR, see Section 2.3.3) when obtaining the variances of the point estimates. It appears that when there is no model misspecification, that the confidence intervals based on the sandwich estimator have similar coverage levels as the confidence intervals based on the design-based estimates. However, when there is model misspecification, the design-based confidence intervals have coverage that is unexpectedly large, implying that the variance estimates are too large.

The third goal of this chapter is to investigate the assertion that adding sampling weights can compensate for model misspecification in LME models. The history of this assertion is summarized in Section 1.2.5. The simulations in this chapter indicate that the weights can help for model misspecification only when the model misspecification induces informative sampling. Bias related to a misspecified model that does not relate to the sampling design are unaffected by the sampling weights.

The fourth goal of this chapter is to investigate the assertion that adding sampling weights can compensate for informative sampling in LME models. The history of this assertion is summarized in Section 1.2.5. The simulations in this chapter support those conclusions. The inverse sampling weights can help compensate for bias induced by informative sampling, though they do not eliminate the bias.

The last goal of this chapter is to investigate the different scalings of the weights, which are introduced in Section 2.4.2. Recall from Section 2.4.1 that the weighted LME estimates are consistent if the number of clusters increases as the population size increases. If the conditional weights ($w_{i|k}$, the inverse probability that individual i is sampled provided

cluster k is in the sample) are multiplied by a cluster level constant, then the consistency argument remains unchanged. This allows us to consider scalings of the weights to reduce the bias in the variance components. Three different scalings are considered in Section 2.4.2 and the effects of the scalings on the bounds of a balanced random intercept model are in Tables 2.1 and 2.2. The simulations in this chapter compare the different scalings when the data are not balanced and when the models are more complicated than random intercept models. These simulations found that the unweighted estimate has the smallest variance. When there is informative sampling, the unweighted estimates are biased. The weighted unscaled estimate corrects the bias in the fixed effects, but produces more bias in the random effects. The scaled 1 weightings remove the bias in the fixed effects, and overcorrect for the weighted unscaled bias in the random effects. The scaled 2 weightings remove the bias in the fixed effects and are in between the weighted unscaled and weighted scaled 1 bias in the random effects.

This chapter also contains a number of appendices collected together in Section 3.7 that provide additional detail about the simulation methods and results. In particular, Section 3.7.6 summarizes the computer code written to run the simulation sand provides web-links to the code for the interested reader.

3.2 Previous LME Simulation Results

3.2.1 Overview

Table 3.1 contains a summary of the previous simulation designs performed by the authors of the methods described in this thesis. As mentioned in Section 2.2.3 the method by RHS was also published concurrently by ASP, whose simulation results are included in Table 3.1. I present the previous simulation studies in the same order as I presented the methods in Chapter 2. This order represents the order of the location in which weights are inserted in the pseudo-maximum likelihood estimation (see Section 2.2). Recall that RHS (and ASP) insert the weights before the derivative is taken, KG insert the weights immediately after the derivative is taken, and PSHGR insert the weights in the process of solving for the parameter values.

In evaluating the previous studies with respect to the goals of this chapter, note that none of the authors compared their method to the other methods presented in this thesis, so there are no previous direct comparisons. All the authors' estimating models matched their generating models, so there was no model misspecification in previous simulations. Below, I summarize the authors' studies based upon the third and fourth goals listed above; to investigate the effect of weights on informative sampling and to compare the different scalings of the weights. In addition to my goals listed above, many of the authors were interested in the effect of sample sizes on the estimates and these are also listed in Table 3.1. Finally, I will also note the authors' methods of computing variances of their point estimates.

		RHS	ASP	KG	PSHGR
	Simulation Comparions	None	None	None	None
Generated (and Estimated) Model	Random Intercept Model: $Y_{ik} = \beta_0 + U_{0k} + \epsilon_{ik}$		✓	✓	✓
	Logistic Random Intercept Model: $\text{logit}(Y_{ik}) = \beta_0 + \beta_1 x_{1ik} + \beta_2 x_{2k} + U_{0k}$	✓	✓		
	Two level model	✓	✓	✓	✓
	Multiple-level model		✓		
Sampling Scheme	Non Informative Cluster, Non Informative Elements		✓		✓
	Informative Cluster, Non Informative Elements		✓	✓	✓
	Non Informative Cluster, Informative Elements		✓	✓	
	Informative Cluster, Informative Elements	✓			✓
	Weights and Scalings [a]	U, WU, WS1, WS2 WS1IS[b], WS2IS[b], Method C[b]	U, WU, WS1, WS2	U, WU, WS1, WS2	U, WU, WS1, WS2
Population and Sample Sizes	Cluster Population Size (K)	500	Unknown	1500	300
	Cluster Sample Size (k)	about 300	100	33, 99	35, 75
	# Population Elements per Cluster (N_k)	5, 10, 20, 50, 100	Unknown	100, 5	Random: 38 to 147
	# Sampled Elements per Cluster (n_k)	0.5 N_k	5, 20, 100	75, 5, 4	9, 38, $0.4N_k$, $0.1N_k$

Table 3.1: Summary of Previous Simulation Study Designs

[a] U = Unweighted, WU = Weighted Unscaled, WS1 = Weighted Scaled 1, WS2 = Weighted Scaled 2

[b] WS1IS = Weighted Scaled 1 Invariant Selection, WS2IS = Weighted Scaled 2 Invariant Selection. See Section 3.2.3 for more details

3.2.2 RHS Simulation Summary

Rabe-Hesketh and Skrondal (2006), denoted RHS, performed simulations with a logistic random intercept model, one cluster level covariate, x_{1k}, and one individual level covariate, x_{2ik},

$$\log\left(\frac{P(Y_{ik}=1)}{1-P(Y_{ik}=1)}\right) = 1 + x_{1k} + x_{2ik} + U_{0k}$$

Their finite population contains 500 clusters, each with the same number of elements per cluster (either 5, 10, 20, 50 or 100). They oversample clusters whose absolute value of the random effect (U_{0k}) was less than one and oversample individuals whose random error (ϵ_{ik}) is less than zero. They sample approximately 300 clusters and approximately half of the elements in the sampled cluster. The RHS results are summarized in Table 3.2. For this table, an estimate was labeled biased if the confidence interval (mean over 100 iterations $\pm$ 2 times standard deviation of the 100 iterations divided by 10) did not contain the true value.

When analyzing the effect of the weights on informative sampling, note that the undersampling of large random intercepts (i.e. undersample $|U_{0k}| \geq 1$) should cause the unweighted estimates of σ^2_{0k} to be too small and the undersampling of error terms greater than zero (i.e. $\epsilon_{ik} \geq 0$) should cause the unweighted estimate of β_0 to have negative bias.

As can be seen from Table 3.2, the unweighted estimate of β_0 is biased under all sample sizes. The weights reduce this bias, however it is not until the cluster population sizes are $N_k = 50$ that the bias becomes negligible (recall from Table 3.1 that the sample size is

roughly half of the population size). The unweighted estimates of σ^2_{0k} are also biased. The effect of adding the weights is mixed for σ^2_{0k}. For the N_k=5, the bias is reduced by all the weights. For the other values of N_k, there is at least one weighting scheme that produces the same (or larger) bias than the unweighted estimate and there are some weighting schemes that appear to do well, however none of the weighting schemes eliminate the bias.

When analyzing the the differences in the scaling of the weights, recall that the scaling is to help correct the bias in the weighted unscaled estimates of the random effect variances. The weighted unscaled estimates of σ^2_{0k} have a positive bias. Both the scaled 1 and scaled 2 estimates appear to overcorrect this bias, resulting in negative bias for the corresponding weighted estimates of σ^2_{0k}, however the bias of the weighted scaled 2 estimates appear to be smaller than the bias in the weighted scaled 1 estimates. For the larger population sizes, $N_k = 20$ or 50, the weighted unscaled estimates do as well or better than the weighted scaled 2 estimates.

RHS use the sandwich estimator to compute the standard errors of the point estimates. To evaluate the variances, RHS simulate the model 1000 times when the cluster size was $N_k = 50$ (while sampling 1/2 of the elements per cluster) and computed confidence intervals. The coverage of the RHS 95% confidence intervals created with the sandwich estimate variances range from 94.1% to 94.7% for the fixed effects, and is 92.4% for σ^2_{0k}.

3.2.3 ASP Simulation Summary

Asparouhov (2006), denoted ASP, performed quite extensive simulations in his paper. These simulations vary the type of informative sampling, the intraclass correlation and

	Design	Weighting Scheme	β_0	β_1	β_2	σ^2_{0k}
Simulation 1	Clusters:	unweighted	bias (0.60)	bias (0.08)	bias (0.06)	bias (0.61)
N_k=5	Undersample $\|U_k\| > 1$	weighted unscaled	unbiased	bias (0.19)	bias (0.22)	bias (0.47)
	Elements:	weighted scaled 1	bias (0.32)	unbiased	bias (0.06)	bias (0.42)
	Undersample $\epsilon_{ik} > 0$	weighted scaled 2	bias (0.25)	unbiased	unbiased	bias (0.30)
Simulation 2	Clusters:	unweighted	bias (0.63)	bias (0.13)	bias (0.14)	bias (0.23)
N_k=10	Undersample $\|U_k\| > 1$	weighted unscaled	bias (0.04)	bias (0.06)	bias (0.11)	bias (0.19)
	Elements:	weighted scaled 1	bias (0.17)	bias (0.09)	bias (0.09)	bias (0.60)
	Undersample $\epsilon_{ik} > 0$	weighted scaled 2	bias (0.12)	bias (0.06)	unbiased	bias (0.26)
Simulation 3	Clusters:	unweighted	bias (0.64)	bias (0.16)	bias (0.16)	bias (0.18)
N_k=20	Undersample $\|U_k\| > 1$	weighted unscaled	unbiased	bias (0.05)	bias (0.05)	bias (0.09)
	Elements:	weighted scaled 1	bias (0.09)	bias (0.06)	bias (0.05)	bias (0.30)
	Undersample $\epsilon_{ik} > 0$	weighted scaled 2	bias (0.06)	unbiased	unbiased	bias (0.17)
Simulation 4	Clusters:	unweighted	bias (0.65)	bias (0.18)	bias (0.18)	bias (0.13)
N_k=50	Undersample $\|U_k\| > 1$	weighted unscaled	unbiased	unbiased	bias (0.02)	bias (0.05)
	Elements:	weighted scaled 1	bias (0.04)	unbiased	bias (0.02)	bias (0.13)
	Undersample $\epsilon_{ik} > 0$	weighted scaled 2	unbiased	unbiased	unbiased	bias (0.06)

Table 3.2: RHS Simulation Design and Results

the model being simulated. He uses many scalings for the weights, including unweighted, weighted unscaled, weighted scaled 1 and weighted scaled 2 estimation methods. ASP ran one simulation comparing the unweighted, weighted unscaled, weighted scaled 1 and weighted scaled 2 weights. He investigated the effect of the intra-class correlation on the weighted scaled 2 estimates and looked at a multilevel logistic regression with weighted scaled 2 estimates. The results of his simulations are summarized in Table 3.3.

For the informative sampling and intra-class correlation simulations, ASP uses the random intercept model,

$$y_{ik} = 0.5 + U_{0k} + \epsilon_{ik}, \quad U_{0k} \sim N(0, 0.5), \quad \epsilon_{ik} \sim N(0, 2) \tag{3.1}$$

where the population cluster size is 100, and the number of sampled individuals per cluster

Design	Recommended Weighting Scheme	Notes
Informative Sampling	Alternate method where all weights are scaled by the estimated population size divided by the sample size	Weighted Scaled 1 and Weighted Scaled 2 both also did well. All methods do best when cluster size is large or informativeness is weak.
Intra-Class Correlation	Weighted Scaled 2	Only Weighted Scaled 2 was analyzed. It was confirmed that when the ICC is small all parameters exhibit more bias.
Multi-Level Logistic	Weighted Scaled 2	Only Weighted Scaled 2 method was analyzed. Bias increases as sample size decrease and informativeness increases.

Table 3.3: ASP Simulation Design and Results

is 5, 20 and 100. The population sizes are unknown. The informative sampling simulation samples individuals proportional to $(1+\exp\{-\frac{y_{ik}}{\alpha}\})^{-1}$, where the level of informativeness is determined by the constant α. With this sampling, larger values of y_{ik} are oversampled, which means that elements with larger random intercepts, U_{0k} and/or larger random errors, ϵ_{ik} are oversampled. I would expect to see that the variances of U_{0k} and ϵ_{ik} to be underestimated, with the variance of U_{0k} to be affected more by the informative sampling.

ASP's results are as expected. None of the weighting methods performed well on all three parameters (the intercept and the variances of U_{0k} and ϵ_{ik}) unless the level of informativeness was small, or the sample size was large (over 100). The weighting methods generally correct for the informative sampling in the fixed effects, however for the random effects it takes sample sizes of 100 to see corrections.

When analyzing the differences in the scalings of the weights, the best weighting to use is not clear. For the informative sampling simulation, weighted scaled 1 , weighted scaled 2 and ASP's method C (where the scaling for the weights is the estimated population size

divided by the sample size, $\sum_{ik} w_{ik} / \sum_k n_k$) all perform equivalently.

ASP uses the sandwich estimator to compute the standard errors of the point estimates. He reported the coverage of the corresponding 95% confidence intervals for all estimates.

Finally, ASP also performs simulations that verify that the bias of the variance components increase as the ICC increases. ASP also estimates a multi-level logistic regression model with a random effect and concludes that the bias increases as the sample size decreases and informativeness increases.

3.2.4 KG Simulation Summary

KG are primarily concerned with method of moment estimators, however for the random intercept model with no covariates, the method of moment estimators match the weighted MLE estimates. They ran simulations using a random intercept model,

$$y_{ik} = 1 + U_{0k} + \epsilon_{ik}, \quad U_{0k} \sim N(0,1), \quad \epsilon_{ik} \sim N(0,1).$$

The simulations contain 1500 population clusters (K), of which 33 or 99 are sampled (k). The population cluster sizes (N_k) are 100 and 5, and sample cluster sizes (n_k) are 75 and 5 ($N_k = 100$), and 4 ($N_k = 5$). The goal of KG's method is to improve the small sample properties of the weighted estimators. The bias from the KG simulations are summarized in Table 3.4. They did not report the estimates of β_0. It is unknown how many simulations are averaged for these means, and the variances of these means were not reported for these simulations.

When analyzing the effect of the weights on informative sampling, their simulations

Design	Sampling	σ^2_{0k}	σ^2_ϵ
Clusters:	$k = 33$ or $99, K = 1500$	0.03	unbiased
Undersample $\lvert U_k \rvert > 0.6745$	$n_k = 75, N_k = 100$	0.01	unbiased
Elements:SRS	$k = 33$ or $99, K = 1500$	0.01	0.01
	$n_k = 5, N_k = 100$	unbiased	unbiased
	$k = 33$ or $99, K = 1500$	0.01	0.01
	$n_k = 4, N_k = 5$	0.01	unbiased
Clusters:	$k = 33$ or $99, K = 1500$	0.01	unbiased
Census	$n_k = 75, N_k = 100$	unbiased	unbiased
Elements: SRS	$k = 33$ or $99, K = 1500$	unbiased	0.01
Undersample $\lvert \epsilon_{ik} \rvert > 0.6745$	$n_k = 5, N_k = 100$	unbiased	unbiased
	$k = 33$ or $99, K = 1500$	unbiased	unbiased
	$n_k = 4, N_k = 5$	0.01	unbiased

Table 3.4: KG Simulation Design and Results

show that the bias is effectively removed with their weighted estimates, even with small sample sizes ($K = 1500, k = 33, N_k = 100, n_k = 5$). This is impressive; however the KG method uses additional information that the other methods do not use (the bivariate, trivariate and quadvariate inclusion weights, $w_{ij|k}, w_{ijl|k}, w_{ijlm|k}$, see Section 2.2.4).

KG did not compare their weights in this simulation to unweighted, weighted unscaled, weighted scaled 1 or weighted scaled 2 estimates.

KG use the jackknife estimator for design based survey sampling (see Section 2.3.2) to estimate the variances of their point estimates. They did not compute the variances for the simulation summarized in Table 3.4.

3.2.5 PSHGR Simulation Summary

Stapleton (2002) and Huang and Hidiroglou (2003) conducted simulation studies using the PSHGR method. Their results are not described here as they support the results from the PSHGR simulation study, which is described next. PSHGR ran three simulation studies

varying the level of informative sampling in a random intercept model with no covariates,

$$y_{ik} = 1 + U_{0k} + \epsilon_{ik}, \quad U_{0k} \sim N(0, 0.2), \quad \epsilon_{ik} \sim N(0, 0.5).$$

For each simulation there were 300 population clusters and 35 were sampled. They also ran simulations where 75 clusters were sampled, though they did not show those results and indicated that the results were similar. The number of population elements per cluster, N_k, was random and bounded between 38 and 147 with a mean of 80. They varied the number of sampled elements, n_k, between 38, $0.4 \times N_k$, 9 and $0.1 \times N_k$. The simulations contained different combinations of informative cluster sampling, non-informative cluster sampling, informative individual sampling and non-informative individual sampling. The simulations and results are in Table 3.5. PSHGR did not provide estimates of variances for all the simulated scenarios. As a rule of thumb, in Table 3.5, I marked an estimate as biased if the average over the iterations deviates more than 10% from the true value.

When analyzing the effect of weights on informative sampling, note that sampling clusters proportional to U_{0k} should introduce bias in the unweighted estimates of β_0 and σ^2_{0k}. Sampling of individuals proportional ϵ_{ik} should introduce bias in the estimate of β_0 and σ^2_ϵ. PSHGR found that when there is informative sampling of clusters, the expected biases appear. The use of the weights compensates for the bias in the estimate of β_0, however the effect of the weights on the estimates of the variance components varies according to the sample size.

When analyzing the differences in the scaling of the weights, PSHGR tentatively rec-

	Design	Weighting Scheme	β_0	σ^2_{0k}	σ^2_ϵ
Simulation 1	Clusters: PPS	Unweighted	biased	varied*	biased
	where Size $= U_k$	Weighted Unscaled	unbiased	varied*	unbiased
	Elements	Weighted Scaled 1	unbiased	varied*	varied*
	Undersample $\epsilon_{ik} > 0$	Weighted Scaled 2	unbiased	varied*	unbiased
Simulation 2	Clusters: PPS	Unweighted	biased	varied*	unbiased
	where Size $= U_k$	Weighted Unscaled	unbiased	varied*	varied*
	Elements: SRS	Weighted Scaled 1	unbiased	unbiased	unbiased
		Weighted Scaled 2	unbiased	unbiased	unbiased
Simulation 3	Clusters: PPS	Unweighted	unbiased	unbiased	unbiased
	where Size $= N_k$	Weighted Unscaled	unbiased	varied*	varied*
	Elements: SRS	Weighted Scaled 1	unbiased	unbiased	unbiased
		Weighted Scaled 2	unbiased	unbiased	unbiased

* bias varied according to sample size

Table 3.5: PSHGR Simulation Design and Results

ommended weighted scaled 2 estimates. The bias of the weighted unscaled estimates varied according to the sampling size for all of the sampling scenarios. The weighted scaled 1 and weighted scaled 2 estimates performed better when there was less informative sampling.

PSHGR estimate the variances of the point estimates with design-based methods. They did not estimate the variances in their simulation study.

3.2.6 Summary

From the initial simulations, KG appears to have the lowest bias. The methods need to be compared using the same simulated conditions to get an accurate comparison. Because KG requires the higher order (i.e. bivariate, trivariate and quadvariate) conditional weights, they use more information than the other methods, which may result in better estimates. In reviewing the RHS, ASP and PSHGR simulations, it is not clear which method provides better results.

RHS and ASP found that the coverage levels of the confidence intervals based on the sandwich estimator were very close to the intended coverage. PSHGR provided simulation estimates of variances based on the design based variance estimator but did not evaluate their performance.

None of the four papers investigating the weights contained simulations with model misspecification.

All of the simulations showed that adding weights to the analysis helped compensate for the bias due to informative sampling. The informative sampling in all of the simulations was directly based on either the value of the random effect, U_{0k}, the random error, ϵ_{ik} or the value of the outcome variable, y_{ik}.

RHS and PSHGR both tentatively recommend the weighted scaled two estimates when there is informative sampling. ASP appears to favor weighted scaled 2 weights, as those are the weights used to evaluate the intra-class correlation and the multi-level logistic regression. RHS, ASP and PSHGR found the bias decreases as the sample size increases for all weighting schemes.

3.3 Format of New Simulation Results

Figure 3.1 contains a sample of the format of the new simulation results presented in this dissertation; it is the first row of Figure 3.4, which appears later in Section 3.4 (as do most of the other tables and equations referred to here). The caption on the figure specifies the name and simulation number which correspond to the columns in the summaries in Tables 3.6 and 3.7. Also included in the caption is the equation number of the generating

equation. For Figure 3.1, a summary of the simulation is in the "Mis Ran 5" column of Table 3.6. The generating model for Figure 3.1 is in Equation 3.10. To the left of the plots is the estimated model for the set of variables in that row. In Figure 3.1, the estimated model is in Equation 3.11.

Each of the panels in Figure 3.1 represents a possible parameter in the estimated model. The parameter name is in bold at the top of the plot. Next to the parameter name (in parenthesis) is the variable associated with that parameter, if applicable. Below the parameter name is the range for the parameter. If there is no range (and no plot) printed, then that parameter was not estimated in this model, such as the σ^2_{0k} parameter in Figure 3.1. The solid vertical line indicates the true value of the parameter as it is in the generating model. The horizontal lines represent the 0.025 to 0.975 empirical quantiles over the simulation replicates for that set of estimates. The circle in the horizontal line represents the average of the estimates. Each plot contains eight horizontal line plots.

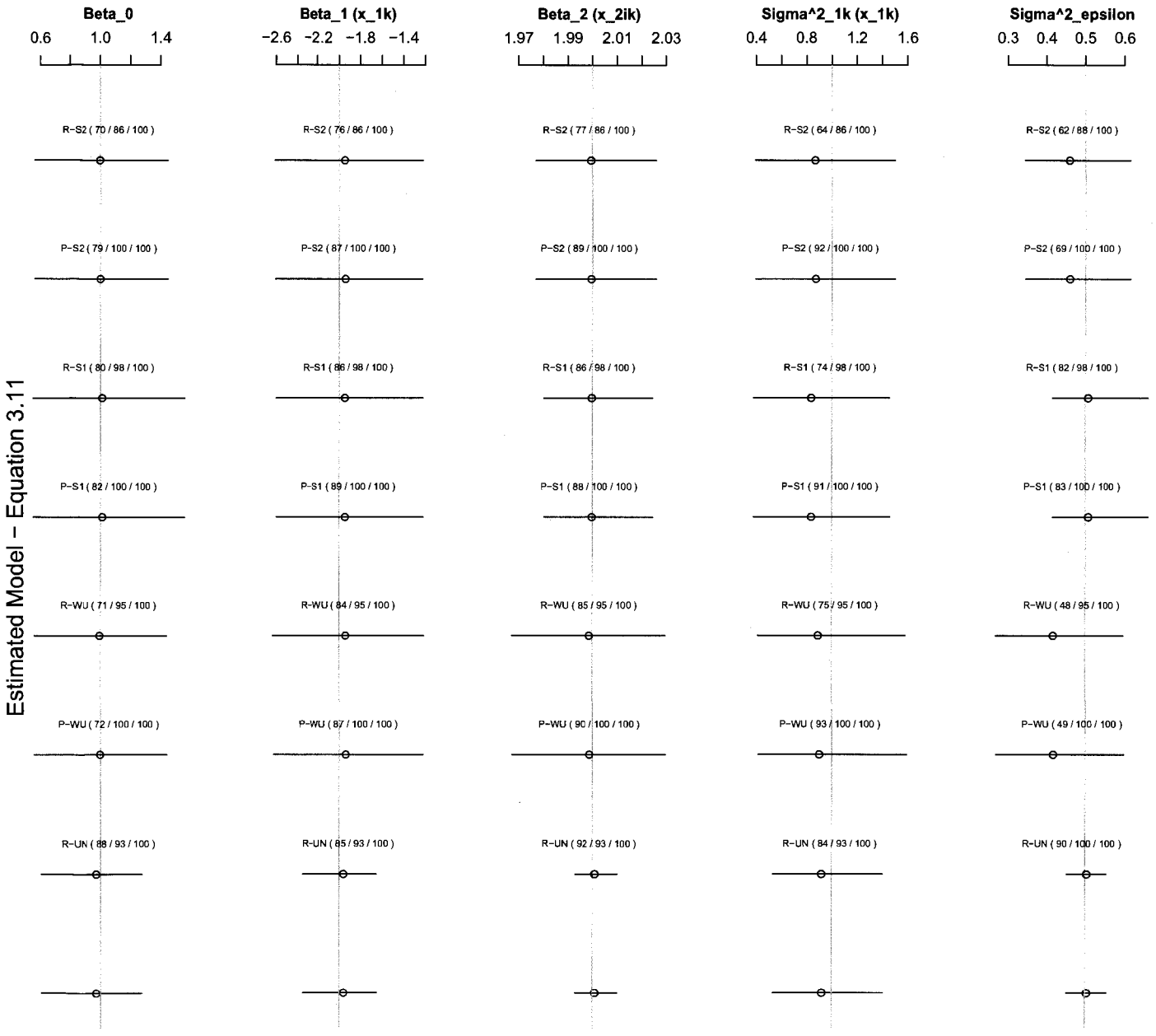

Figure 3.1: Results for Misspecification of Random Variables, Simulation Set 5
Generated Model - Equation 3.10
Sample Presentation Result

The red horizontal line plots represent the PSHGR simulations, and the black horizontal line plots represent the RHS simulations. Each of the horizontal line plots has a caption, such as "R - S2 (70/86/100)". The first term in the caption is either an R (for RHS) or a P (for PSHGR) representing the estimation method used. The second term represents the type of weighting estimation used, either S2 for weighted scaled 2 estimates, S1 for weighted scaled 1 estimates, WU for weighted unscaled estimates or UN for unweighted estimates. Finally, there are three numbers listed. The first number is the number of estimated confidence intervals that contained the true parameter value. These confidence intervals are computed as the point estimate for the given simulation plus or minus 2 times the standard error. The standard error is computed as explained in Section 2.3, where RHS uses the sandwich estimator to estimate the variance and PSHGR use a design based estimate of the variance. The second number represents the number of the iterations where the variance was able to be computed. For RHS, the code to run the simulations is not always able to estimate variances for the estimated point estimates (the estimate of the Hessian is sometimes numerically unstable). Finally, the third number in the caption is the number of iterations where the point estimates are able to be estimated. If the number of iterations is less than 100 for RHS, then it means that some iterations did not converge for any of number of quadrature points between 15 and 35. If the number of iterations is less than 100 for PSHGR, then it means that the iterative generalized least squares algorithm did not converge within 500 iterations.

For example, the RHS weighted unscaled estimates of σ_ϵ^2 are in the fifth plot from the left. The caption on the horizontal line plot is "R - WU (48/95/100)". This means that of

the 100 iterations, all 100 of them are able to produce weighted unscaled estimates of the σ_ϵ^2 parameter. Of the 100 iterations that are able to produce point estimates, 95 of them are able to produce estimates of the variances. Of the 95 iterations able to produce estimates of the variance, 48 of the estimated confidence intervals contained the true parameter value of 0.5. Thus, the estimated coverage of the 95% RHS confidence intervals (as computed with the sandwich estimator variance) is 48/95=50.5%. The horizontal line represent the 0.025 and 0.975 quantiles of the 100 point estimates generated. The average of the 100 estimates is about 0.4, representing a bias of approximately 0.1. When comparing this to the RHS unweighted estimate of σ_ϵ^2, it is clear that there is a smaller spread for the unweighted variances than the weighted variances. In addition, the unweighted estimates are approximately unbiased, and the 95% confidence interval covers the true parameter values 90/100=90% of the time.

3.4 New Simulation Results

The new simulations presented below confirm and expand upon the previously published results. The new simulations that are performed refer to the RHS method, published concurrently with ASP, as the RHS method because the software used to run the simulations was written by Rabe-Hesketh and Skrondal (see `www.gllamm.org`). In addition, the simulations by KG are summarized in this chapter, however I did not perform further simulations of their method as most analysts will not have the joint and quadruple conditional weights ($w_{ij|k}$ and $w_{ijlm|k}$) weights needed to implement their method.

There are a total of 12 simulation sets, broken into 4 categories: 1) Misspecification

		Mis Fix[a] 1	Mis Fix[a] 2	Mis Fix[a] 3	Mis Fix[a] 4	Mis Ran[b] 5	Mis Ran[b] 6	Mis Ran[b] 7	Mis Ran[b] 8
Generating Model	Random Intercept Model: $Y_{ik} = 1 + U_{0k} - 2x_{1k} + 2x_{2ik} + \epsilon_{ik}$, $U_{0k} \sim N(0, 0.2)$, $\epsilon_{ik} \sim N(0, 0.5)$	✓	✓	✓	✓				
	Random Slope Model: $Y_{ik} = 1 + (-2 + U_{1k})x_{1k} + 2x_{2ik} + \epsilon_{ik}$, $U_{0k} \sim N(0, 0.2), \epsilon_{ik} \sim N(0, 0.5)$					✓	✓		
	Random Slope Model: $Y_{ik} = 1 + -2x_{1k} + (2 + U_{2k})x_{2ik} + \epsilon_{ik}$, $U_{0k} \sim N(0, 0.2), \epsilon_{ik} \sim N(0, 0.5)$							✓	✓
Estimated Model	Same as generated	✓	✓	✓	✓	✓	✓	✓	✓
	Missing x_{1k}	✓	✓	✓	✓				
	Missing x_{2ik}	✓	✓	✓	✓				
	Missing U_{1k} added random intercept U_{0k}					✓	✓		
	Missing U_{2k} added random intercept U_{0k}							✓	✓
Sampling Scheme	Cluster Sample PPS N_k, element sampling PPS independent variable	✓				✓		✓	
	Cluster Sample PPS N_k, element sampling PPS x_{2ik}		✓						
	Cluster Sample PPS x_{1k}, element sampling PPS independent variable			✓					
	Cluster Sample PPS x_{1k}, element sampling PPS x_{2ik}				✓				
	Cluster Sample PPS U_{1k}, element sampling PPS independent variable						✓		
	Cluster Sample PPS U_{2k}, element sampling PPS independent variable								✓

Table 3.6: Simulation Designs for the Misspecification of Fixed and Random Effects

[a]Mis Fix = Misspecification of Fixed Effects
[b]Mis Ran = Misspecification of Random Effects

		Mis Strat[c] 9	Mis Strat[c] 10	Mis Strat[c] 11	Mis Clust[d] 12
Generated Model	Generated Model: Random Intercept with necessary adjustments reflecting the sampling design	✓	✓	✓	✓
Estimated Model	Estimated Model: Random Intercept with necessary adjustments reflecting the sampling design	✓	✓	✓	✓
Generated Layers	Stratified / Clustered	✓			
	Clustered / Stratified		✓		
	Stratified / Clustered / Stratified			✓	
	Cluster 1/ Cluster 2				✓
Estimated Layers	Stratified / Clustered	✓		✓	
	Clustered / Stratified		✓	✓	
	Clustered	✓	✓	✓	
	Clustered 1				✓
	Clustered 2				✓
Sampling Scheme	Clusters Sampling PPS Size, Element Sampling PPS independent variable	✓	✓	✓	✓
	Clusters Sampling PPS U, Element Sampling PPS independent variable	✓	✓		

Table 3.7: Simulation Designs for the Misspecification of Stratification and Clustering Layers

[c]Mis Strat=Misspecification of Stratification Layers
[d]Mis Clust= Misspecification of Clustering Layers

of the Fixed Effects, 2) Misspecification of the Random Effects, 3) Misspecification of Stratification Layers and 4) Misspecification of Clustering Layers. The simulation sets are summarized in Tables 3.6 and 3.7. Each simulation category contains model misspecification and/or informative sampling. Recall the definitions of sampling completely at random, sampling at random, sampling not at random (or informative sampling) from Section 1.2, as they are used to describe the extent of informative sampling in the simulations.

The summary of each simulation reflects on the conclusions from this chapter.

1. The PSHGR and RHS point estimates are very similar. The differences in the point estimates are due to numerical instabilities in the estimation procedures.

2. The sandwich estimator, used by RHS , is a better estimator of the variance of the point estimates than the design-based variance estimator used by PSHGR. However, the sandwich estimator is not as numerically stable since computation of the Hessian is not always possible. The PSHGR design-based variance estimator appears reasonable when the model is correctly specified, however the estimates are sometimes too large when the model is misspecified, especially for the variance components.

3. When there is model misspecification that does not induce informative sampling, weighted estimates do not reduce bias of the estimators.

4. When there is informative sampling, the weighted estimators do reduce the bias of the point estimates, though they do not eliminate it.

5. The unweighted estimate has the smallest variance. When there is informative sampling, the unweighted estimates are biased. The weighted unscaled estimate corrects

the bias in the fixed effects, but produces bias in the random effects. The scaled 1 weightings remove the bias in the fixed effects, and usually reduces (or overcorrects) for the weighted unscaled bias in the random effects. The scaled 2 weightings remove the bias in the fixed effects and are in between the weighted unscaled and weighted scaled 1 bias in the random effects. There are some cases where the scaled 1 estimates are more biased in the same direction as the weighted unscaled estimates. In these cases, the weighted scaled 2 estimates are still between the weighted unscaled and weighted scaled 1 weights. The variation of the estimates across the 100 iterations are somtimes similar for all estimates (weighted or unweighted). When the variation across the 100 iterations varies by the weighting, then the smallest variation is in the unweighted estimates, followed by the weighted scaled 1, weighted scaled 2 and unweighted estimates.

3.4.1 Misspecification of Fixed Effects - Non-Informative Sampling - Simulation Set 1

A summary of this simulation set is in the "Mis Fix 1" column of Table 3.6. The generating model is a random intercept model,

$$y_{ik} = 1 + U_{0k} - 2x_{1k} + 2x_{2ik} + \epsilon_{ik}, \quad U_{0k} \sim N(0, 0.2), \quad \epsilon_{ik} \sim N(0, 0.5), \qquad (3.2)$$

where $x_{1k} \sim N(3, 9)$ and $x_{2ik} \sim N(1, 25)$. There are 300 population clusters, with a random uniform number of population units per population cluster between 50 and 100. The sample contains 35 clusters and 20 units per cluster. The sampling of clusters is

proportional to the magnitude of the population cluster size, N_k. Sampling of individuals within clusters is proportional to an independently generated random variable assigned to each element[1]. There are three estimated models in this simulation set. One matches the generated model, one removes the fixed effect for x_{1k} , and one removes the fixed effect for x_{2ik},

$$y_{ik} = \beta_0 + U_{0k} + \beta_1 x_{1k} + \beta_2 x_{2ik} + \epsilon_{ik}, \; U_{0k} \sim N(0, \sigma^2_{0k}), \; \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \quad (3.3)$$

$$y_{ik} = \beta_0 + U_{0k} + \beta_2 x_{2ik} + \epsilon_{ik}, \; U_{0k} \sim N(0, \sigma^2_{0k}), \; \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \quad (3.4)$$

$$y_{ik} = \beta_0 + U_{0k} + \beta_1 x_{1k} + \epsilon_{ik}, \; U_{0k} \sim N(0, \sigma^2_{0k}), \; \epsilon_{ik} \sim N(0, \sigma^2_\epsilon). \quad (3.5)$$

The sampling scheme is sampling completely at random for all three estimated models.

Summary

The results from this simulation set are in Figure 3.2. A detailed description of the results is in Section 3.7.1.

In this simulation, the estimation using the PSHGR method generally matched the estimation using the RHS method. Some differences between PSHGR and RHS appear in Figure 3.2. The PSHGR unweighted estimates of σ^2_{0k} from Equation 3.4 have a larger mean and a larger 0.025 empirical quantile than RHS.

[1]Each element was assigned a random variable $a_{ik} \sim \text{Uniform}(-5, 5)$. They were then sampled proportional to $(1 + \exp(-a_{ik}))^{-1}$.

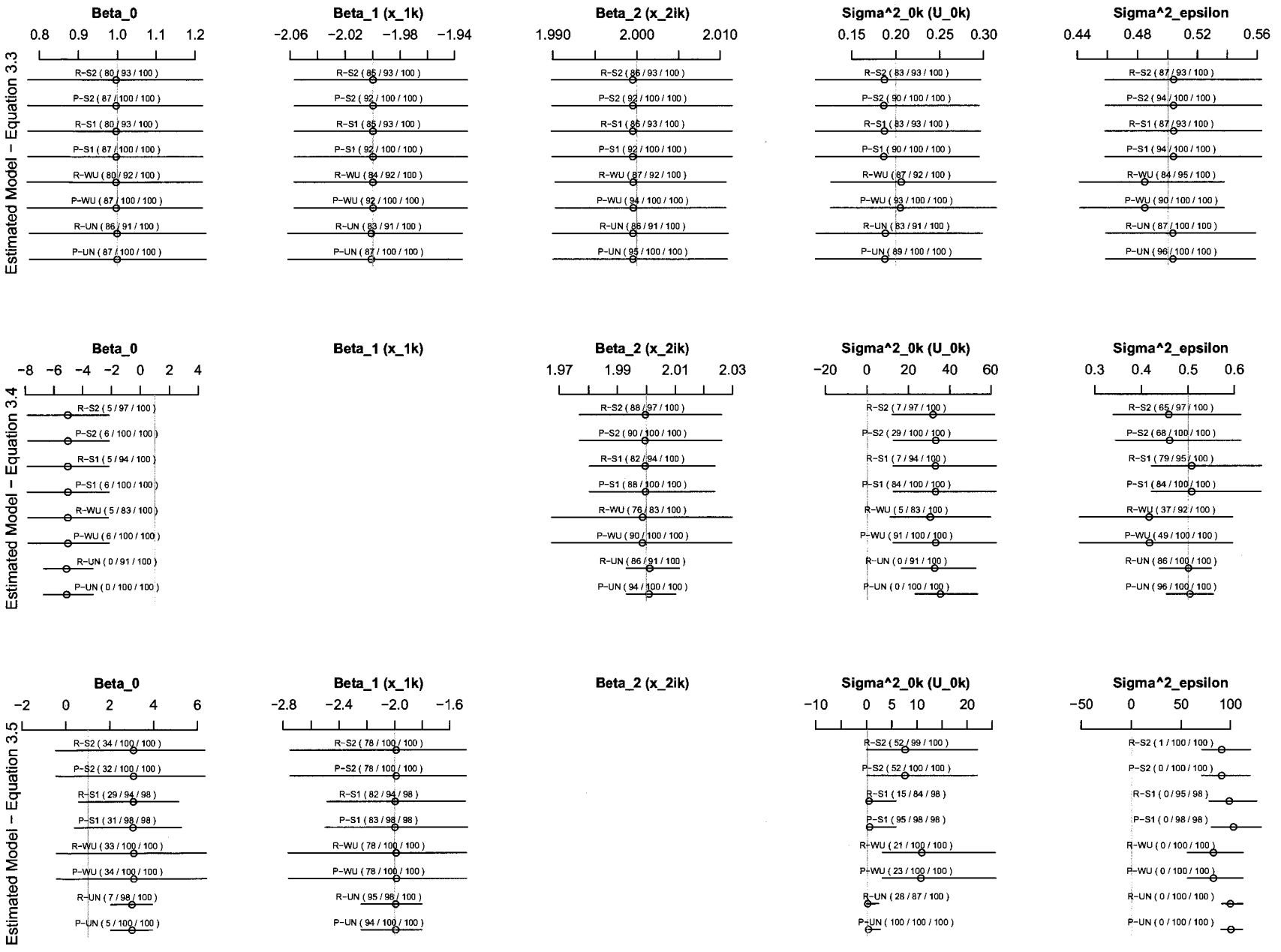

Figure 3.2: Results for Misspecification of Fixed Effects, Simulation Set 1
Generated Model - Equation 3.2

The PSHGR weighted unscaled estimate of σ^2_{0k} from Equation 3.4 has a larger mean and larger 0.025 and 0.975 empirical quantiles than RHS. Finally, the PSHGR weighted scaled 1 estimate of σ^2_ϵ has a larger mean and larger 0.025 and 0.975 quantiles than RHS. These differences (and smaller differences not visible in Figure 3.2) are due to numerical instabilities in the RHS and PSHGR estimations, and are described in detail in Section 3.7.1.

When analyzing the coverage of the confidence intervals, look at the simulation where the estimating model is from Equation 3.3, which matches the generating model and has the least bias. The coverage from the RHS 95% confidence interval coverage varies between 85% and 95% in the fixed effects and between 83% and 87% in the variance components. For PSHGR, the 95% confidence interval coverage varies between 87% and 95% for the fixed effects and between 89% to 96% in the variance components. RHS produced sandwich estimates for the variance for between 83 and 100 of the 100 iterations for the estimates in Figre 3.2. The PSHGR estimates of the variance of σ^2_{0k} were quite large in estimated models from Equations 3.4 and 3.5, causing the confidence interval coverage to be much larger than the coverage from RHS. This may indicate a problem with the variance estimator for PSHGR. To verify this, the coverage of the confidence intervals for the expected parameter value should be obtained.

The second and third estimated models from Equations 3.4 and 3.5 contained model misspecification. When a covariate was included in the generating model but not the estimating model, a model misspecification bias was found in all weighting methods. The removal of a fixed covariate caused the intercept to change by the mean of the missing

covariate times its associated parameter. The variance of the missing covariate moved into the intercept variance (if it was a cluster covariate) or the random error variance (if it was in individual covariate). It is possible that the missing covariate could affect both variance estimates if the covariate was an individual covariate whose mean varied across clusters. See Section 3.7.1 for more details for this simulation. The various weighting methods did not help against model misspecification bias.

These simulations did not contain any informative sampling, so there was no informative sampling bias.

All weighting methods provided similar mean estimates of the β coefficients. The 0.024 and 0.975 quantiles over the simulation runs sometimes vary according to weighting scheme. When the model is correctly specified, all estimates (weighted and unweighted) have similar spread across the simulations. When the model is misspecified, the spreads sometimes differ. When they do, the unweighted has the smallest spread, followed by the weighted scaled 1, weighted scaled 2 and weighted unscaled estimates. There is a difference in the weighting schemes with the estimation of the variance components. The weighted unscaled estimates have a bias, the weighted scaled 1 estimate compensates (or overcompensates) for the weighted unscaled bias and the weighted scaled 2 bias is between the weighted scaled 1 and the weighted unscaled bias. How close the weighted scaled 2 bias is to the weighted scaled 1 bias appears to vary. When the model is correctly specified, the weighted scaled 1 and weighted scaled 2 estimates of the variance components (both σ^2_ϵ and σ^2_{0k}) are close. When there is model misspecification, the weighted scaled 2 estimates appear to be balanced in between the weighted scaled 1 and the weighted unscaled estimates.

3.4.2 Misspecification of Fixed Effects - Partially Informative Sampling - Simulation Sets 2 and 3

A summary of these simulations sets are in the "Mis Fix 2" and "Mis Fix 3" columns of Table 3.6. The generating model for both simulation sets is a random intercept model,

$$y_{ik} = 1 + U_{0k} - 2x_{1k} + 2x_{2ik} + \epsilon_{ik} \quad U_{0k} \sim N(0, 0.2), \quad \epsilon_{ik} \sim N(0, 0.5).$$

where $x_{1k} \sim N(3, 9)$ and $x_{2ik} \sim N(1, 25)$. The population has 300 clusters, each with a random number of units per cluster between 50 and 100. The sample contains 35 clusters and 20 units per cluster. The three estimated models are

$$\begin{aligned} y_{ik} &= \beta_0 + U_{0k} + \beta_1 x_{1k} + \beta_2 x_{2ik} + \epsilon_{ik}, \quad U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \\ y_{ik} &= \beta_0 + U_{0k} + \beta_2 x_{2ik} + \epsilon_{ik}, \quad U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \\ y_{ik} &= \beta_0 + U_{0k} + \beta_1 x_{1k} + \epsilon_{ik}, \quad U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon). \end{aligned}$$

Result Description for Misspecification of Fixed Effects – Simulation Set 2

For Misspecification of Fixed Effects - Simulation Set 2, the sampling of clusters is proportional to the magnitude of the population cluster size, N_k. The sampling of individuals in a cluster is proportional to the magnitude of the individual level covariate x_{2ik}. The simulation is sampling at random when the covariate x_{2ik} is included in the estimating model, and informative sampling when the estimating model did not contain the covariate. When the model did contain the covariate x_{2ik}, then the estimation behaved exactly as in

Misspecification of Fixed Effects - Simulation Set 1, where there is no informative sampling. When the model did not contain the x_{2ik} covariate, the estimation behaved exactly as in Misspecification of Fixed Effects - Simulation Set 4, where there is informative sampling. For space considerations, the results for Misspecification of Fixed Effects - Simulation Set 2 were not presented here.

Result Description for Misspecification of Fixed Effects - Simulation Set 3

For Misspecification of Fixed Effects - Simulation Set 3, the sampling of clusters is proportional to the magnitude of the cluster level covariate x_{1k}. The sampling of individuals is proportional to an independently generated random variable assigned to each element[2]. The simulation was sampling at random when the variable x_{1k} was included in the estimating model, and informative sampling when the estimating model did not contain the covariate x_{1k}. When the model did contain the covariate x_{1k}, then the estimation behaved exactly as in Misspecification of Fixed Effects - Simulation Set 1, where there is no informative sampling. When the model did not contain the x_{1k} covariate, the estimation behaved exactly as in Misspecification of Fixed Effects - Simulation Set 4, where there is informative sampling. For space considerations, the results for Misspecification of Fixed Effects - Simulation Set 3 were not presented here.

[2]Each element was assigned a random variable $a_{ik} \sim \text{Uniform}(-5, 5)$. They were then sampled proportional to $(1 + \exp(-a_{ik}))^{-1}$.

3.4.3 Misspecification of Fixed Effects - Informative Sampling - Simulation Set 4

A summary of this simulation set is in the "Mis Fix 4" column of Table 3.6. The generating model is a random intercept model,

$$y_{ik} = 1 + U_{0k} - 2x_{1k} + 2x_{2ik} + \epsilon_{ik} \quad U_{0k} \sim N(0, 0.2), \quad \epsilon_{ik} \sim N(0, 0.5), \tag{3.6}$$

where $x_k \sim N(3, 9)$ and $x_{ik} \sim N(1, 25)$. There are 300 population clusters, with a random uniform number of population units per population cluster between 50 and 100. The sample contains 35 clusters and 20 units per cluster. The sampling of clusters is proportional to the magnitude of the cluster covariate, x_{1k}. Sampling of individuals is proportional to the magnitude of the individual covariate, x_{2ik}. The three estimated models are

$$y_{ik} = \beta_0 + U_{0k} + \beta_1 x_{1k} + \beta_2 x_{2ik} + \epsilon_{ik}, \quad U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \tag{3.7}$$

$$y_{ik} = \beta_0 + U_{0k} + \beta_2 x_{2ik} + \epsilon_{ik}, \quad U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \tag{3.8}$$

$$y_{ik} = \beta_0 + U_{0k} + \beta_1 x_{1k} + \epsilon_{ik}, \quad U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \tag{3.9}$$

The simulation is sampling at random when both the x_{1k} and x_{2ik} covariates are included in the estimating model, and informative sampling when the estimating model does not contain either (or both) of the covariates.

Summary

The results from this simulation set are in Figure 3.3. A detailed description of the results is in Section 3.7.1.

In this simulation, the estimation using the PSHGR method mostly matched the estimation using the RHS method. There are some differences between the PSHGR and RHS estimates, but they are not large enough to be seen in Figure 3.3. See Section 3.7.1 for more details.

The coverage of the confidence intervals between RHS and PSHGR are similar for the estimated model in Equation 3.7. The RHS 95% confidence intervals for the β coefficients are between 73% and 97% and for the variance components they are between 58% to 87% . The coverage of the PSHGR 95% confidence intervals for the β coefficients from the estimated model in Equation 3.7 are between 76% and 97% and for the variance components they are between 56% and 88%. The major difference between PSHGR and RHS is that PSHGR can compute the variances of the point estimates in all the simulation runs for all the parameters, whereas RHS computes the variances between 87% and 100% of the simulation runs. In addition, the PSHGR confidence intervals for σ^2_{0k} in the estimated model from Equation 3.8 have larger coverage than expected, especially for the weighted unscaled estimate. This may indicate a problem with the variance computation for PSHGR. To verify this, the coverage of the confidence intervals for the expected parameter value should be obtained.

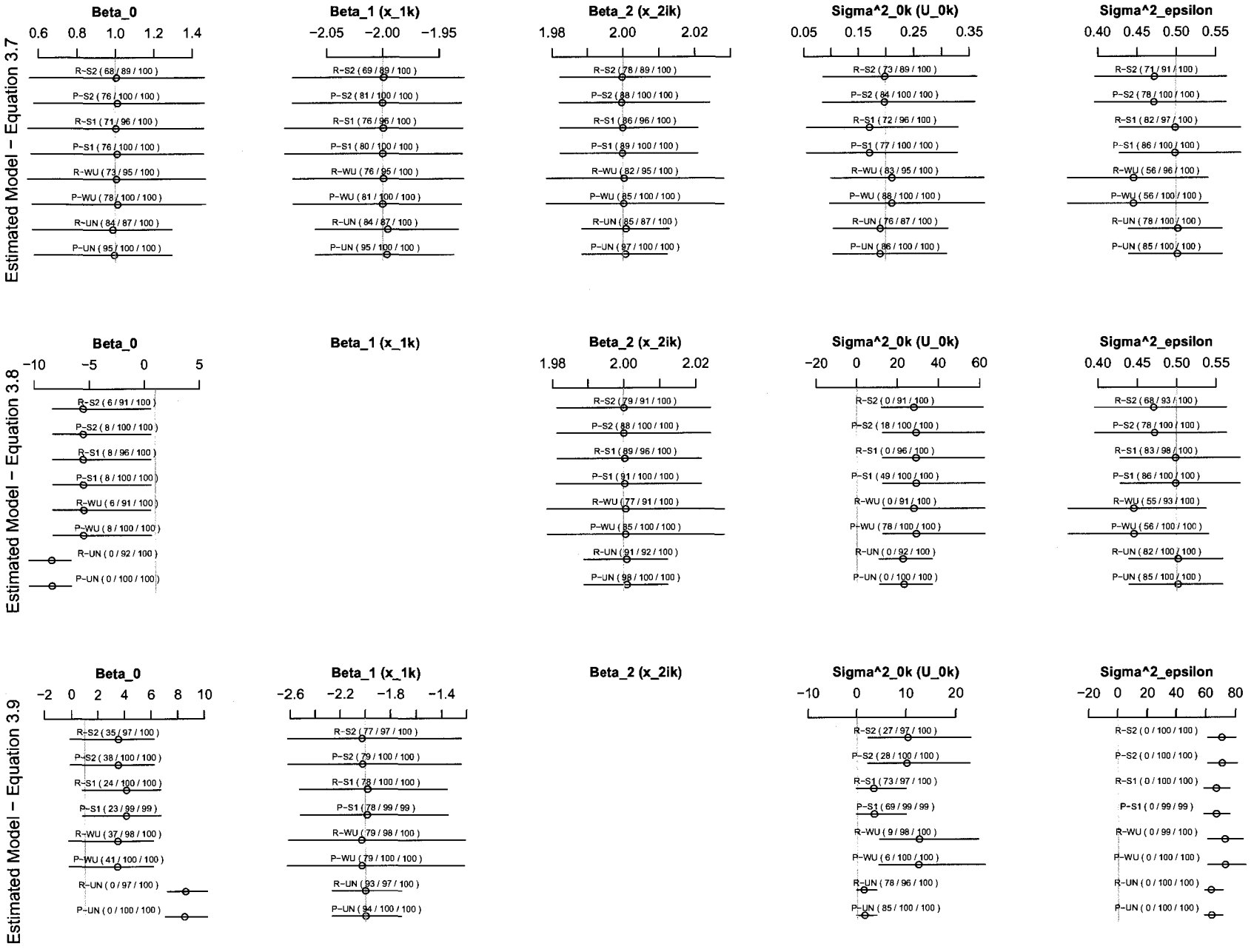

Figure 3.3: Results for Misspecification of Fixed Effects, Simulation Set 4
Generated Model - Equation 3.6

The second and third estimated models contain model misspecification. The estimated models from Equations 3.8 and 3.9 contain model misspecification that induces informative sampling. For the estimated model defined in Equation 3.8, the weighted estimates of the intercept are near -5, as the are in Figure 3.2 under the estimated model in Equation 3.4 where there is no informative sampling. This is not near the true value of 1. This difference in the estimates is due to the model misspecification that is not related to informative sampling. A similar trend is seen in the estimate of the intercept from the estimated model in Equation 3.9. The weighted methods do not compensate for the model misspecification bias.

The second and third estimated models contain informative sampling. The informative sampling bias can be seen by comparing the unweighted estimates to the weighted estimates for β_0 from estimating models in Equations 3.8 and 3.9. It can also be seen in the estimates for σ^2_{0k}, but it is not so obvious. The unweighted estimate of σ^2_{0k} is the same size or larger than the weighted unscaled estimate of σ^2_{0k} in Figure 3.2 under the estimating model in Equation 3.4 where there is no informative sampling. However, the unweighted estimate of σ^2_{0k} is smaller than the weighted unscaled estimate of σ^2_{0k} in Figure 3.3 under the estimating model in Equation 3.8 where there is informative sampling. The same can be seen under estimated models in Equations 3.5 and 3.9, however it is not so clear since these estimates are against the constraint that $\sigma^2_{0k} \geq 0$. See section 3.7.1 for more details. Note that the weights do not fully compensate for the informative sampling bias, as can be seen by comparing the estimates of σ^2_ϵ from the estimated model in Equation 3.9 to the estimates of σ^2_ϵ in Figure 3.2 under the estimated model in Equation 3.5. The addition of the weights

helped to compensate for the informative sampling.

All weighting methods generally provide similar point estimates and ranges for the β coefficients. The exception is that the spread of the weighted unscaled estimates of β_2 appear to be larger. The estimates for β_0 from the estimated model in Equation 3.9 vary more then the other β coefficients. For both σ^2_ϵ and σ^2_{0k}, there is some bias in the unweighted estimates. The weighted unscaled estimates have a larger bias, the weighted scaled 1 estimate compensates (or overcompensates) for the weighted unscaled bias and the weighted scaled 2 bias is in between the weighted scaled 1 and the weighted unscaled bias. Note that the weighted scaled 2 estimates of σ^2_ϵ and σ^2_{0k} when the model is correctly specified are further from the weighted scaled 1 estimates than in Figure 3.2. This indicates that the scaled 2 weights may help with estimation of the variance components under sampling at random. The unweighted estimates a smaller 0.975, 0.025 quantile spread than the weighted estimates in all these simulations. When the spreads of the weighted estimates vary, then the weighted unscaled spread is the largest, followed by the weighted scaled 2 estimates spread and the weighted scaled 1 estimates spread.

3.4.4 Misspecification of Random Variables - Non-Informative Sampling - Simulation Set 5

A summary of this simulation set is in the "Mis Ran 5" column of Table 3.6. The generating model is a random slope model with the random slope on a cluster level covariate,

$$y_{ik} = 1 + (-2 + U_{1k})x_k + 2x_{ik} + \epsilon_{ik} \quad U_{1k} \sim N(0,1), \quad \epsilon_{ik} \sim N(0, 0.5), \qquad (3.10)$$

where $x_{1k} \sim N(3,9)$ and $x_{2ik} \sim N(1,25)$. There are 300 population clusters, with a random uniform number of population units per population cluster between 50 and 100. The sample contains 35 clusters and 20 units per cluster. The sampling of clusters is proportional to the magnitude of the population cluster size, N_k. Sampling of individuals within clusters is proportional to an independently generated random variable assigned to each element[3]. There are two estimated models in this simulation set. One matches the generated model, and one removes the random slope U_{1k} and adds a random intercept U_{0k},

$$y_{ik} = \beta_0 + (\beta_1 + U_{1k})x_k + \beta_2 x_{ik} + \epsilon_{ik}, \quad U_{1k} \sim N(0, \sigma^2_{1k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \qquad (3.11)$$

$$y_{ik} = \beta_0 + \beta_1 x_k + \beta_2 x_{ik} + U_{0k} + \epsilon_{ik}, \quad U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \qquad (3.12)$$

The sampling scheme is sampling competely at random for both estimated models.

[3]Each element was assigned a random variable $a_{ik} \sim$ Uniform$(-5, 5)$. They were then sampled proportional to $(1 + \exp(-a_{ik}))^{-1}$.

Summary

The results from this simulation set are in Figure 3.4. A detailed description of the results is in Section 3.7.1.

In this simulation, the estimation using the PSHGR method mostly matched the estimation using the RHS method. There are some differences between the PSHGR and RHS estimates, but they are not large enough to be seen in Figure 3.4. See Section 3.7.1 for more details.

The coverage of the confidence intervals of RHS and PSHGR are similar, with the RHS 95% confidence intervals for the β coefficients from the estimated model in Equation 3.11 are between 75% to 95%, and for the variance components they are between 50% and 90%. The coverage of the PSHGR 95% confidence intervals for the β coefficients from the estimated model in Equation 3.11 are between 72% to 95%, and for the variance components they are between 49% and 96%. RHS was able to produce sandwich estimator variances for between 77% and 100% of the simulation runs, while PSHGR was able to produce design based estimator variances for 100% of the simulation runs. Again, the number of confidence intervals for PSHGR covering the true parameter appears larger than the RHS intervals, especially for the random effects and for the estimated model in Equation 3.12, where there is model misspecification. This may indicate a problem with the variance estimator for PSHGR. To verify this, the coverage of the confidence intervals for the expected parameter value should be obtained.

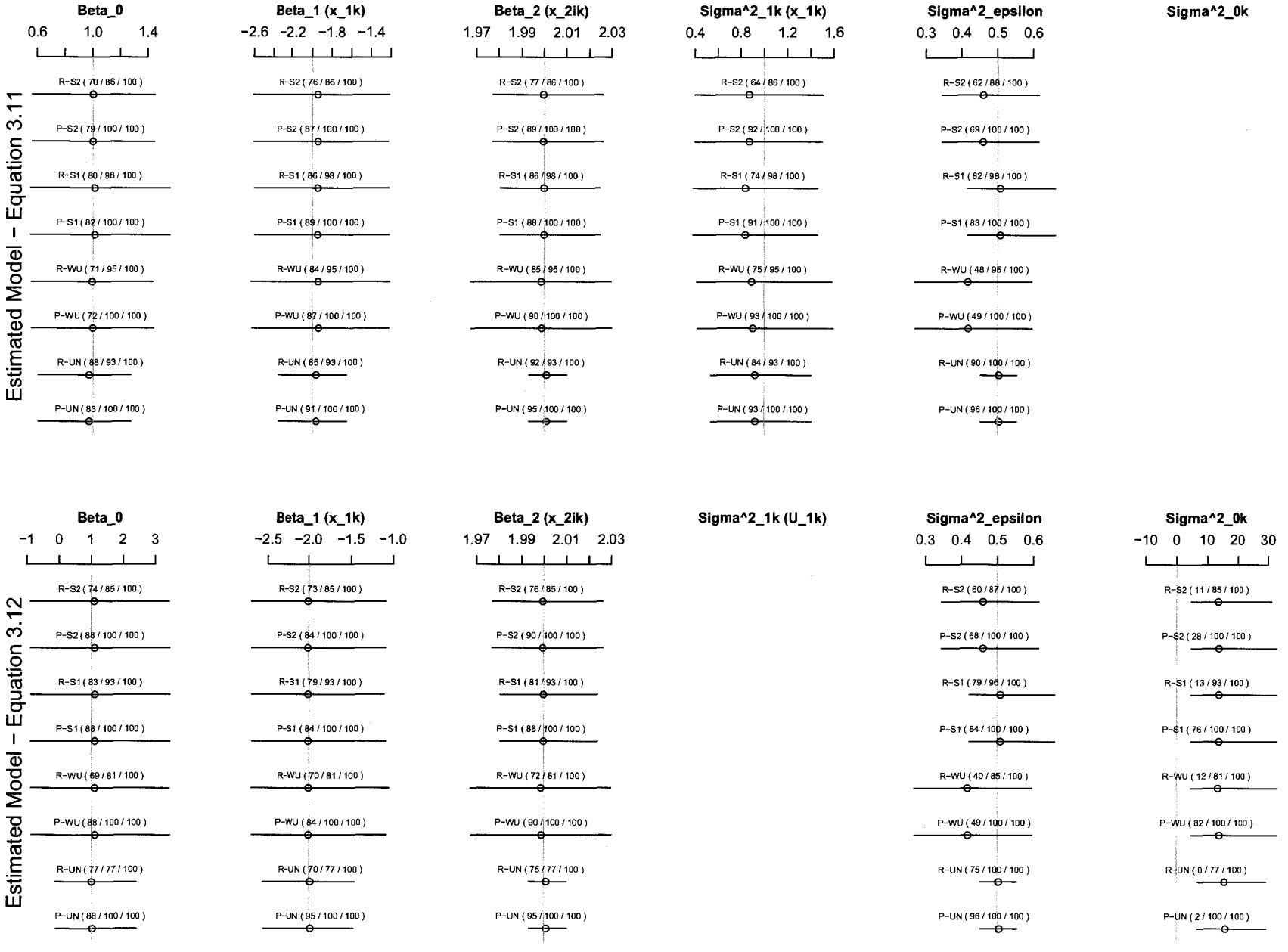

Figure 3.4: Results for Misspecification of Random Variables, Simulation Set 5
Generated Model - Equation 3.10

The second estimated model contains model misspecification. The random slope term is removed and a random intercept term is added. The random intercept variance contains the variance of the random slope term ($U_{1k}x_{1k}$), however there is some negative bias in the estimates. The expected variance of the random intercept is approximately 18, while the simulated means are between 13.5 and 16, see Section 3.7.1 for details. This is expected due to the low intra-class correlation, see Section 2.4.2. None of the other estimates are affected by the model misspecification. Note that the weighted estimates do not appear to compensate for the model misspecification, though it is not entirely clear what compensating for model misspecification would mean in this example.

These simulations did not contain any informative sampling, so there was no informative sampling bias.

All weighting schemes provide similar point estimates and ranges for the β parameters. The exception is that the spread for the weighted unscaled estimate of β_2 is larger than the other weighted schemes. The variance of the unweighted estimates is smaller. The estimates of the random slope follow the trend that the weighted scaled 2 estimate is between the weighted unscaled and the weighted scaled 1. The bias doesn't quite follow the same pattern as the weighted scaled 1 estimates show more bias in the same direction as the weighted unscaled, as opposed to σ^2_ϵ and σ^2_{0k} where the weighted scaled 1 compensates for the bias in the weighted unscaled estimates. All the unweighted estimates a smaller 0.975, 0.025 quantile spread than the weighted estimates. When the spreads of the weighted estimates vary, then the weighted unscaled spread is the largest, followed by the weighted scaled 2 estimates spread and the weighted scaled 1 estimates spread.

3.4.5 Misspecification of Random Variables - Informative Sampling - Simulation Set 6

A summary of this simulation set is in the "Mis Ran 6" column of Table 3.6. The generating model is a random slope model, with the random slope on the cluster level covariate,

$$y_{ik} = 1 + (-2 + U_{1k})x_{1k} + 2x_{2ik} + \epsilon_{ik} \quad U_{1k} \sim N(0,1), \quad \epsilon_{ik} \sim N(0, 0.5), \quad (3.13)$$

where $x_{1k} \sim N(3,9)$ and $x_{2ik} \sim N(1,25)$. There are 300 population clusters, with a random number of population units per population cluster between 50 and 100. The sample contains 35 clusters and 20 units per cluster. The sampling of clusters is proportional to the magnitude of the random effect, U_{1k}. Sampling of individuals within clusters is proportional to an independently generated random variable assigned to each element[4]. There are two estimated models in this simulation set. One matches the generated model, and one removes the random slope U_{1k} and adds a random intercept U_{0k},

$$y_{ik} = \beta_0 + (\beta_1 + U_{1k})x_k + \beta_2 x_{2ik} + \epsilon_{ik}, \quad U_{1k} \sim N(0, \sigma^2_{1k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \quad (3.14)$$

$$y_{ik} = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2ik} + U_{0k} + \epsilon_{ik}, \quad U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \quad (3.15)$$

The sampling scheme is informative sampling at the cluster level.

[4]Each element was assigned a random variable $a_{ik} \sim$ Uniform$(-5, 5)$. They were then sampled proportional to $(1 + \exp(-a_{ik}))^{-1}$.

Results Summary

The results from this simulation set are in Figure 3.5. A detailed description of the results is in Section 3.7.1.

In this simulation, the estimation using the PSHGR method mostly matched the estimation using the RHS method. The PSHGR estimate of β_0 under the estimated model in Equation 3.15 has a lower mean and a lower 0.025 quantile and a higher 0.975 quantile than the corresponding RHS estimate. This and other differences between PSHGR and RHS are described in more detail in Section 3.7.1. The coverage of the confidence intervals of RHS and PSHGR are similar, with the RHS 95% confidence intervals for the β coefficients from the estimated model in Equation 3.14 are between 10% to 96%, and for the variance components they are between 31% and 89%. The coverage of the PSHGR 95% confidence intervals for the β coefficients from the estimated model in Equation 3.14 are between 11% to 95%, and for the variance components they are between 41% and 94%. RHS was able to produce sandwich estimator variances for between 80% and 100% of the simulation runs, while PSHGR was able to produce design based estimator variances for 100% of the simulation runs. In general, the number of PSHGR confidence intervals that cover the true parameter is larger than for RHS, especially when the model is misspecified as in the estimated model in Equation 3.15. This may indicate a problem with the variance computation for PSHGR. To verify this, the coverage of the confidence intervals for the expected parameter value should be obtained.

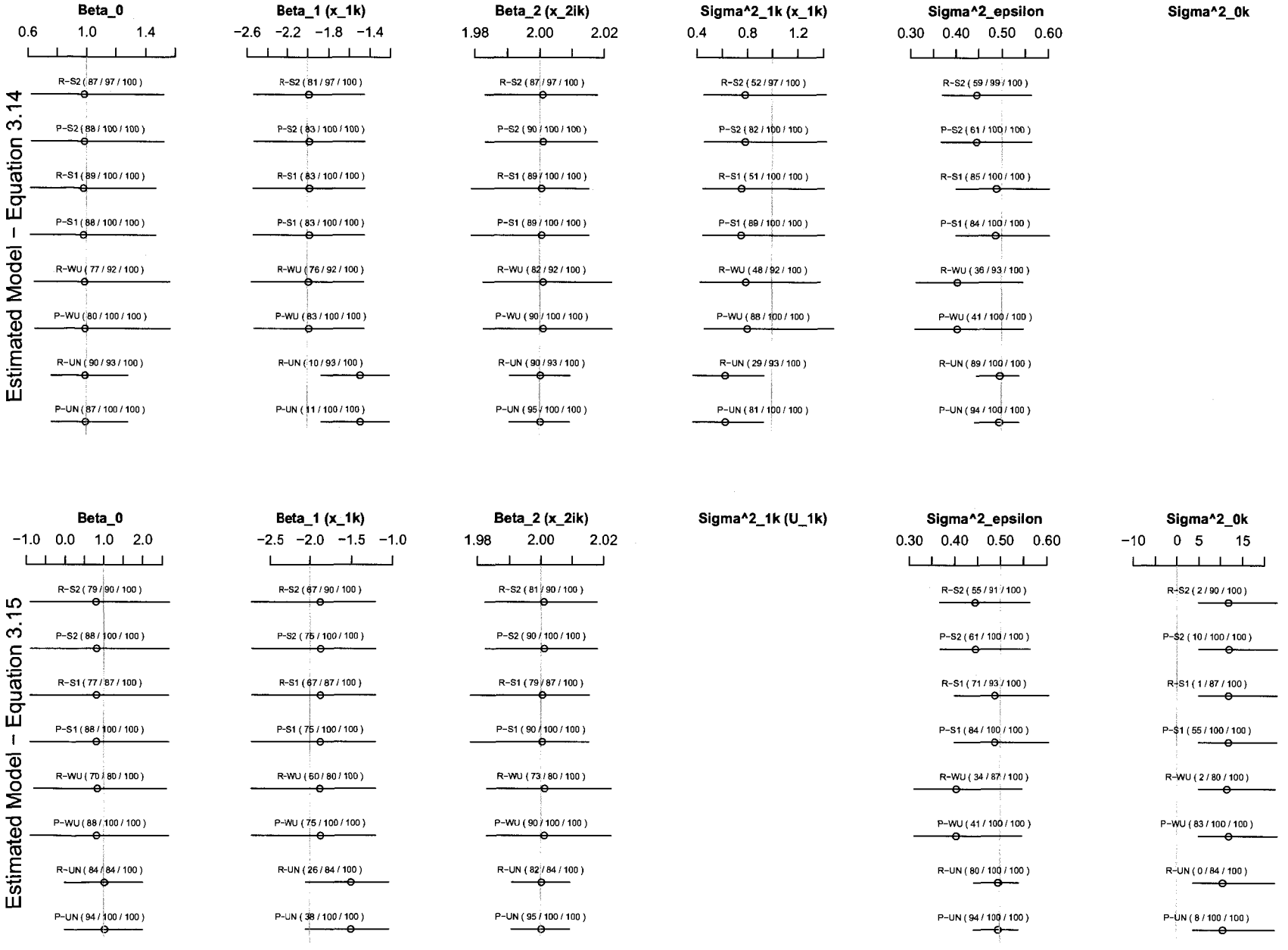

Figure 3.5: Results for Misspecification of Random Variables, Simulation Set 6
Generated Model - Equation 3.13

The second estimated model contained model misspecification. The random slope term was removed and a random intercept term was added. The random intercept variance contained the variance fo the random slope term ($U_{1k}x_{ik}$). None of the other estimates were affected by the model misspecification.

Both estimated models contain informative sampling. When the estimated and generated models match each other, the informative sampling causes the unweighted estimates of β_{1k} and σ^2_{1k} to be biased. All of the weighted estimates help to compensated for this informative sampling. When the random slope is removed from the model and a random intercept is added, the estimate of β_1 contained the same informative sampling bias in the unweighted estimate. The informative sampling bias of the σ^2_{1k} estimate is now reflected in the estimate of σ^2_{0k}. When comparing the unweighted estimate of σ^2_{0k} to the same estimate from the estimating model from Equation 3.12, it is clear that the unweighted estimate from the estimating model in Equation 3.15 is smaller. None of the other terms were affected.

All the weighted estimates performed similarly for the β coefficients. As in the previous simulations, for σ^2_ϵ and σ^2_{0k}, the weighted unscaled estimates are biased, the weighted scaled 1 estimates overcompensate for the bias, and the weighted scaled 2 estimates are in between. Note that unlike the previous simulation set that was non-informative, the pattern of the weights in the estimate of σ^2_{1k} follows the pattern of the other variance components. The unweighted estimates a smaller 0.975, 0.025 quantile spread than the weighted estimates in all these simulations. When the spreads of the weighted estimates vary, then the weighted unscaled spread is the largest, followed by the weighted scaled 2 estimates spread and the

weighted scaled 1 estimates spread.

3.4.6 Misspecification of Random Variables - Non-Informative Sampling - Simulation Set 7

A sumary of this simulation set is in the "Mis Ran 7" column of Table 3.6. The generating model is a random slope model, where the random slope is on the individual level covariate,

$$y_{ik} = 1 - 2x_{1k} + (2 + U_{2k})x_{2ik} + \epsilon_{ik} \;\; U_{2k} \sim N(0, 0.8), \;\; \epsilon_{ik} \sim N(0, 0.5). \quad (3.16)$$

where $x_{1k} \sim N(3, 9)$ and $x_{2ik} \sim N(1, 25)$. There are 300 population clusters, with a random uniform number of population units per population cluster between 50 and 100. The sample contains 35 clusters and 20 units per cluster. The sampling of Clusters is proportional to the magnitude of the population cluster size, N_k. Sampling of individuals within clusters is proportional to an independently generated random variable assigned to each element[5]. There are two estimated equations in this simulation set. One matches the generated model, and one removes the random slope U_{2k} and adds a random intercept U_{0k},

$$y_{ik} = \beta_0 + U_{1k}x_{1k} + (\beta_2 + U_{2k})x_{2ik} + \epsilon_{ik}, \;\; U_{2k} \sim N(0, \sigma^2_{2k}), \;\; \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \quad (3.17)$$

$$y_{ik} = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2ik} + U_{0k} + \epsilon_{ik}, \;\; U_{0k} \sim N(0, \sigma^2_{0k}), \;\; \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \quad (3.18)$$

This sampling scheme is sampling completely at random.

[5]Each element was assigned a random variable $a_{ik} \sim$ Uniform$(-5, 5)$. They were then sampled proportional to $(1 + \exp(-a_{ik}))^{-1}$.

Summary

The results from this simulation set are in Figure 3.6. A detailed description of the results is in Section 3.7.1.

In this simulation, the estimation using the PSHGR method mostly matched the estimation using the RHS method. There are some differences between the PSHGR and RHS estimates, but they are not large enough to be seen in Figure 3.6. See Section 3.7.1 for more details. The coverage of the confidence intervals of RHS and PSHGR are mostly similar, with the RHS 95% confidence intervals for the β coefficients from the estimated model in Equation 3.20 are between 77% to 95%, and for the variance components they are between 49% and 94%. The coverage of the PSHGR 95% confidence intervals for the β coefficients from the estimated model in Equation 3.17 are between 78% to 92%, and for the variance components they are between 59% and 98%. Note that the coverage of the σ^2_{2k} estimates for PSHGR (approximately 85/100) is much higher than the estimates of the coverage for RHS (approximately 45/95). The RHS coverages appear more accurate given the bias in the estimates. This may indicate a problem with the variance estimator for PSHGR. To verify this, the coverage of the confidence intervals for the expected parameter value should be obtained. RHS was able to produce sandwich estimator variances for between 92% and 100% of the simulation runs, while PSHGR was able to produce design based estimator variances for 100% of the simulation runs. In addition, for the estimated model in Equation 3.18, a simulation run did not converge for the RHS weighted scaled 2 estimates. The number of confidence intervals for PSHGR covering the true value fo σ^2_{2k} under the estimated model in Equation 3.17 are larger than the corresponding RHS

intervals.

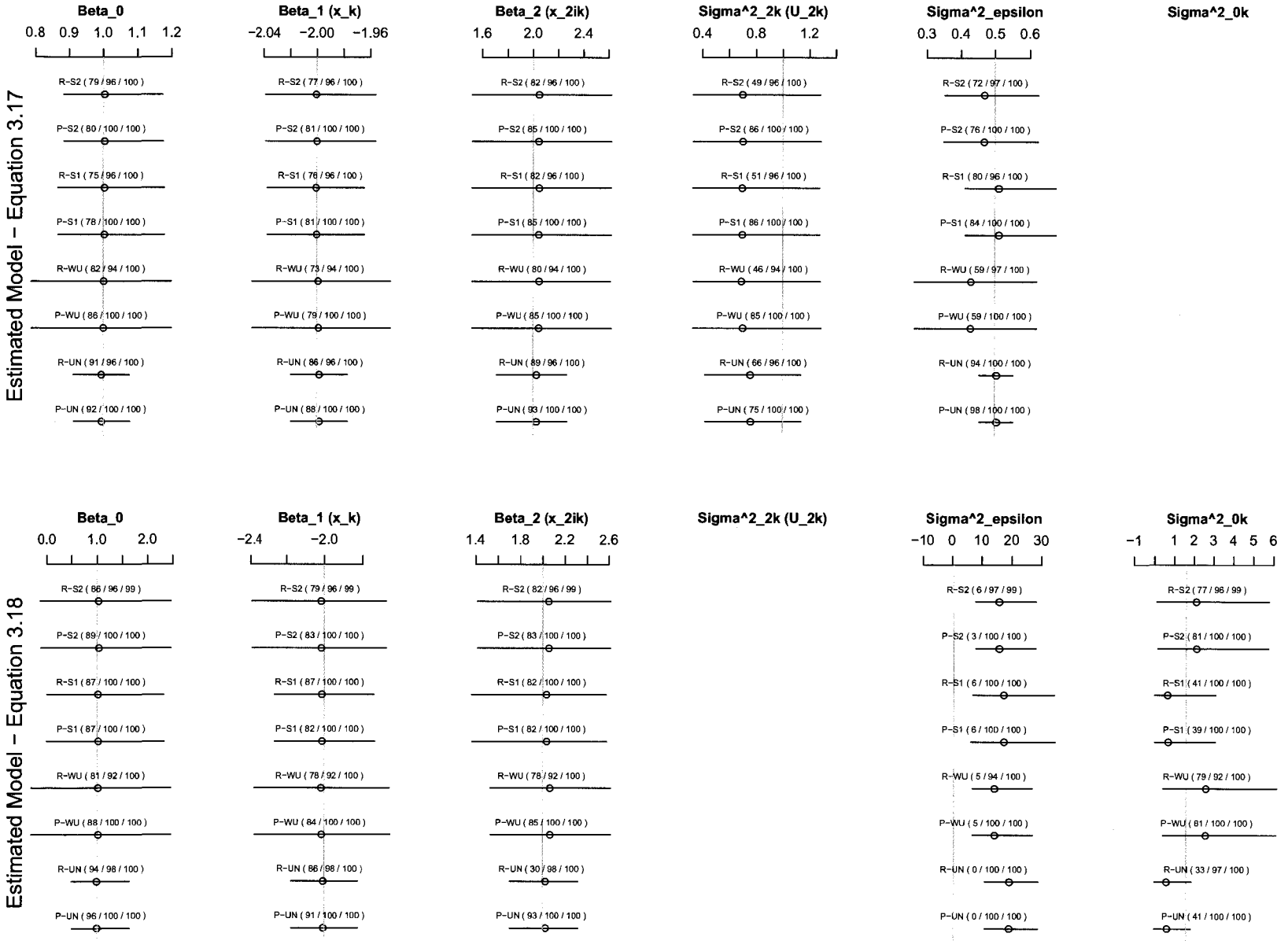

Figure 3.6: Results for Misspecification of Random Variables, Simulation Set 7
Generated Model - Equation 3.16

The second estimated model contains model misspecification. The variance from the dropped random slope is split between the estimated variance of the random intercept and the estimated variance of the random error, as expected from the description in Section 3.7.1. The estimates of β are not affected by the model misspecification. The addition of the weights does not help compensate for this model misspecification.

These simulations does not contain any informative sampling, so there is no informative sampling bias.

All the weighting schemes perform equivalently for the β estimates, except the weighted estimates of β_0 and β_1 with the unscaled weights have slightly larger variances. The weighting of the variance components follows the trend that the weighted unscaled estimates are biased, the weighted scaled 1 overcompensates for the bias, and the weighted scaled 2 estimates are between the weighted scaled 1 and the weighted unscaled estimates. An exception to this is the estimate of σ^2_ϵ for the estimated model in Equation 3.18. Here we see that the weighted unscaled estimates are biased, and that the weighted scaled 1 estimates are more biased than the weighted scaled 1, with the weighted scaled two still between the weighted scaled 1 and the unweighted estimates. The unweighted estimates a smaller 0.975, 0.025 quantile spread than the weighted estimates in all these simulations. When the spreads of the weighted estimates vary, then the weighted unscaled spread is the largest, followed by the weighted scaled 2 estimates spread and the weighted scaled 1 estimates spread. The exception is in the estimated model in Equation 3.18 for the estimates of β_2 and σ^2_ϵ, where the scaled 1 estimates simulation spread is larger than the weighted scaled 2 spread.

3.4.7 Misspecification of Random Variables - Informative Sampling - Simulation Set 8

A summary of this simulation set is in the "Mis Ran 8" column of 3.6. The generating model is a random slope model, with the random slope on a cluster level covariate,

$$y_{ik} = 1 - 2x_{1k} + (2 + U_{2k})x_{2ik} + \epsilon_{ik} \;\; U_{2k} \sim N(0, 0.8), \;\; \epsilon_{ik} \sim N(0, 0.5), \quad (3.19)$$

where $x_{1k} \sim N(3, 9)$ and $x_{2ik} \sim N(1, 25)$. There are 300 population clusters, with a random uniform number of population units per population cluster between 50 and 100. The sample contains 35 clusters and 20 units per clusters. The sampling of clusters was proportional to the magnitude of the random effect U_{2k}. Sampling of individuals within clusters is proportional to an independently generated random variable assigned to each element[6]. There are two estimated equations in this simulation set. One matches the generated model, and one removes the random slope U_{2k} and adds a random intercept U_{0k},

$$y_{ik} = \beta_0 + U_{1k}x_{1k} + (\beta_2 + U_{2k})x_{2ik} + \epsilon_{ik}, \;\; U_{2k} \sim N(0, \sigma^2_{2k}), \;\; \epsilon_{ik} \sim N(0, \sigma^2_\epsilon) \quad (3.20)$$

$$y_{ik} = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2ik} + U_{0k} + \epsilon_{ik}, \;\; U_{0k} \sim N(0, \sigma^2_{0k}), \;\; \epsilon_{ik} \sim N(0, \sigma^2_\epsilon). \quad (3.21)$$

The sampling scheme is informative sampling for both estimated models.

[6]Each element was assigned a random variable $a_{ik} \sim$ Uniform$(-5, 5)$. They were then sampled proportional to $(1 + \exp(-a_{ik}))^{-1}$.

Summary

The results from this simulation set are in Figure 3.7. A detailed description of the results is in Section 3.7.1.

In this simulation, the estimation using the PSHGR method matched well the estimation using the RHS method. There are no differences to highlight.

The coverage of the confidence intervals of RHS and PSHGR are mostly similar, with the RHS 95% confidence intervals for the β coefficients from the estimated model in Equation 3.20 are between 84% to 94%, and for the variance components they are between 28% and 89%. The coverage of the PSHGR 95% confidence intervals for the β coefficients from the estimated model in Equation 3.20 are between 82% to 95%, and for the variance components they are between 51% and 93%. The number of confidence intervals for PSHGR covering the true parameter appears lager than the RHS intervals, especially for the σ^2_{2k} parameter from the estimated model in Equation 3.20. This may indicate a problem with the variance estimator for PSHGR. To verify this, the coverage of the confidence intervals for the expected parameter value should be obtained. RHS was able to produce sandwich estimator variances for between 95% and 100% of the simulation runs, while PSHGR was able to produce design based estimator variances for 100% of the simulation runs. In addition, for the estimated model in Equation 3.21, there was one simulation for each of the the weighted scaled 2, unweighted and weighted scaled 1 estimates that did not converge for RHS after incrementing the number of quadrature points from 15 to 31.

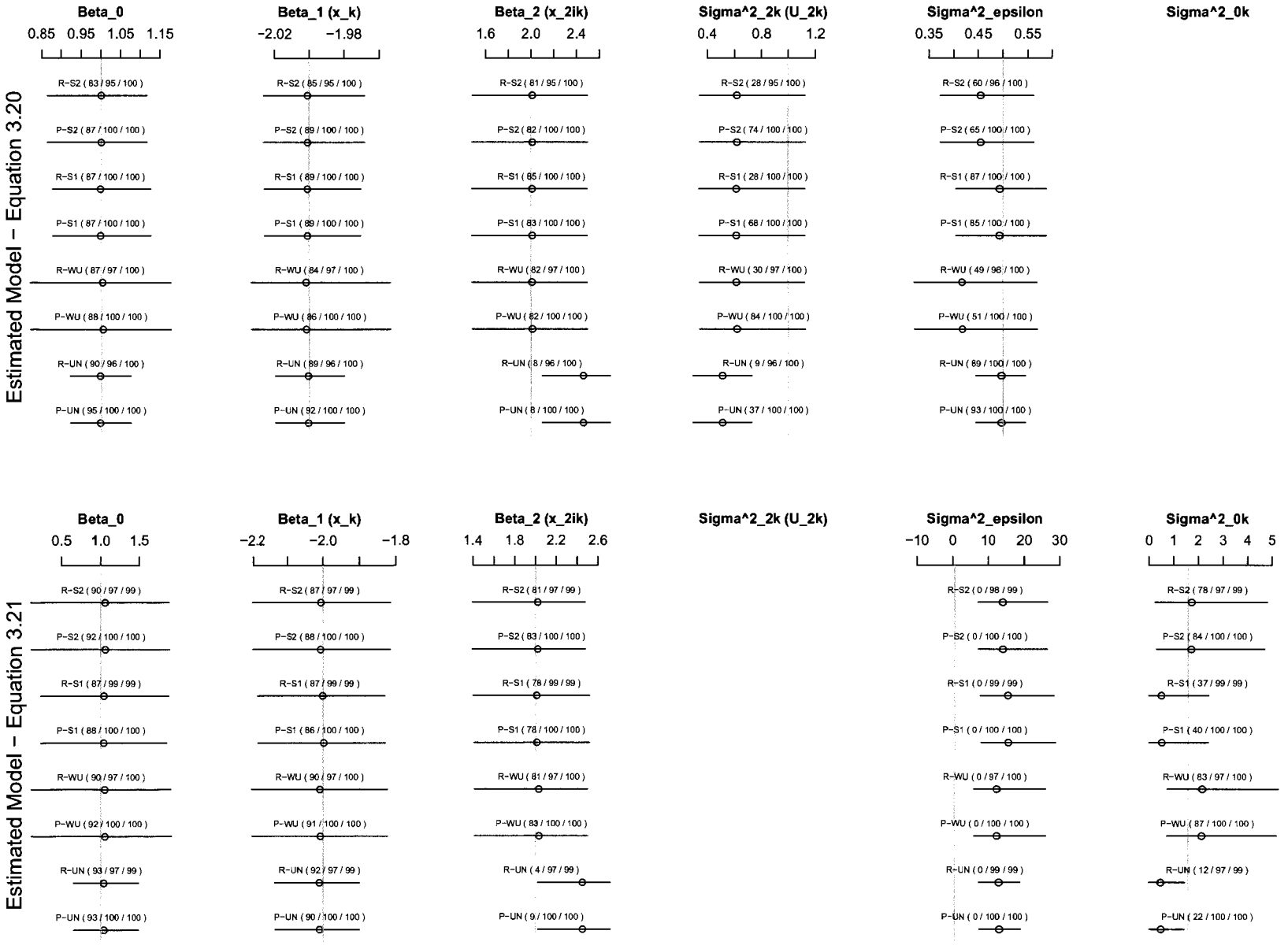

Figure 3.7: Results for Misspecification of Random Variables, Simulation Set 8
Generated Model - Equation 3.19

The second estimated model contains model misspecification. The variance from the dropped random slop is split between the estimated variance of the random intercept and the estimated variance fo the random error, as is expected from the description in Section 3.7.1. The estimates of β were not affected by the model misspecification. The addition of the weights does not help compensate for this model misspecification.

Both estimated models contain informative sampling, the effects of which can be seen in the unweighted estimates of the β_{2ik}, σ^2_{2k} and σ^2_{0k} parameters. In the first estimated model, the unweighted estimate of β_{2ik} is larger than the weighted estimates, and the unweighted estimate of σ^2_{2k} is smaller than the weighted estimates due to oversampling larger values of U_{2k}. In the estimated model from Equation 3.21, the effect of the informative sampling on the β_{2ik} is the same as in Equation 3.20. In addition, the unweighted estimate of σ^2_{0k} is biased low, which can be seen when comparing it to the unweighted estimate of σ^2_{0k} from Equation 3.18 that does not contain the informative sampling.

All of the weighted estimates performed similarly for the β coefficients, however the variance for the weighted unscaled estimates is larger. The pattern in the variance components still holds, the weighted unscaled estimates are biased, the weighted scaled 1 estimates overcompensates for the bias and the weighted scaled 2 estimates are between the weighted unscaled and weighted scaled 1 estimates. The exception to this are the estimates of σ^2_ϵ for the estimated model in Equation 3.21, where the scaled 1 estimates provide more bias in the same direction as the weighted unscaled estimates. The weighted scaled 2 estimates are still between the unweighted and the weighted scaled 1 estimates. The unweighted estimates a smaller 0.975, 0.025 quantile spread than the weighted estimates in all these

simulations. When the spreads of the weighted estimates vary, then the weighted unscaled spread is the largest, followed by the weighted scaled 2 estimates spread and the weighted scaled 1 estimates spread.

3.4.8 Misspecification of Stratification Layers - Stratified / Clustered Sampling - Simulation Set 9

A summary of this simulation set is in the "Mis Strat 9" column of Table 3.7. Let there be two strata where $I_{h==1}(I_{h==2})$ is an indicator variable that the element is in the first (second) stratum, respectively. Within each stratum, there is a layer of clustering. The generating model is a clustered/stratified model,

$$y_{ihk} = -3 + 8I_{h==1} + U_{01k}I_{h==1} + U_{02k}I_{h==2} + \epsilon_{ihk} \tag{3.22}$$

$$U_{01k} \sim N(0,1), U_{02k} \sim N(0,5), \epsilon_{ihk} \sim N(0,0.5), \mathrm{Cov}(U_{01k}, U_{02k}) = 0.$$

This model allows the variance of the clusters in the first stratum to be different from the variance of the clusters in the second stratum. Within each of the two strata, there are 30 population clusters, with a random uniform number of population elements per population cluster between 50 and 100 units. The sample includes 5 clusters from each stratum, and 20 units from each cluster. Sampling of clusters within a stratum is proportional to an independently generated random variable assigned to each cluster[7]. Sampling of elements within a cluster is proportional to an independently generated random variable assigned to each element[8].

There are two estimated models in this simulation set. One matches the generated

[7]Each cluster was assigned a random variable $a_k \sim$ Uniform$(-5,5)$. They were then sampled proportional to $(1+\exp(-a_k))^{-1}$.

[8]Each element was assigned a random variable $a_{ik} \sim$ Uniform$(-5,5)$. They were then sampled proportional to $(1+\exp(-a_{ik}))^{-1}$.

model, and one removes the layer of stratification to estimate a cluster only scheme,

$$\begin{aligned} y_{ihk} &= \beta_0 + \beta_1 I_{h==1} + U_{01k} I_{h==1} + U_{02k} I_{h==2} + \epsilon_{ihk}, \qquad (3.23) \\ & U_{01k} \sim N(0, \sigma^2_{01}), U_{02k} \sim N(0, \sigma^2_{02}), \epsilon_{ihk} \sim N(0, \sigma^2_{\epsilon}), \mathrm{Cov}(U_{01k}, U_{02k}) = \sigma^2_{01k.02k}, \\ y_{ihk} &= \beta_0 + U_{0k} + \epsilon_{ihk} \\ & U_{0k} \sim N(0, \sigma^2_{0k}), \epsilon_{ihk} \sim N(0, \sigma^2_{\epsilon}). \qquad (3.24) \end{aligned}$$

The sampling scheme is at random for both estimated models

These results are presented with the results of an additional simulation. This simulation used the same generating model, but uses informative sampling for the clusters. The generating model is

$$\begin{aligned} y_{ihk} &= -3 + 8 I_{h==1} + U_{01k} I_{h==1} + U_{02k} I_{h==2} + \epsilon_{ihk} \qquad (3.25) \\ & U_{01k} \sim N(0, 1), U_{02k} \sim N(0, 5), \epsilon_{ihk} \sim N(0, 0.5), \mathrm{Cov}(U_{01k}, U_{02k}) = 0, \end{aligned}$$

and there was one estimating equation,

$$\begin{aligned} y_{ihk} &= \beta_0 + U_{0k} + \epsilon_{ihk} \qquad (3.26) \\ & U_{0k} \sim N(0, \sigma^2_{0k}), \epsilon_{ihk} \sim N(0, \sigma^2_{\epsilon}). \end{aligned}$$

Sampling of clusters within a stratum is proportional to the magnitude of the random effect, U_{01k} or U_{02k}, assigned to each cluster. Sampling of elements within a cluster is

proportional to an independently generated random variable assigned to each element[9]. All the other components of the sampling scheme are the same as described after Equation 3.24.

The sampling scheme is informative sampling.

Results Summary

The results from this simulation set are in Figure 3.8. A detailed description of the results is in Section 3.7.1.

In this simulation, the estimation using the PSHGR method mostly matched the estimation using the RHS method. The differences that are visible in Figure 3.8 include all the estimates of $\sigma^2_{01k.02k}$ when the estimated model is in Equation 3.23. This difference is due to the very small point estimates (and no variance) in the PSHGR estimates. In the same estimated model, the PSHGR weighted scaled 1 estimates of σ^2_{02k} have a much larger 0.975 empirical quantile than RHS. In addition, the PSHGR weighted unscaled estimates of σ^2_{0k} from the estimated model in Equation 3.24 has a larger 0.025 and 0.975 empirical quantile than RHS. These and other differences not large enough to be seen in Figure 3.8 are in Section 3.7.1.

[9]Each element was assigned a random variable $a_{ik} \sim \text{Uniform}(-5, 5)$. They were then sampled proportional to $(1 + \exp(-a_{ik}))^{-1}$.

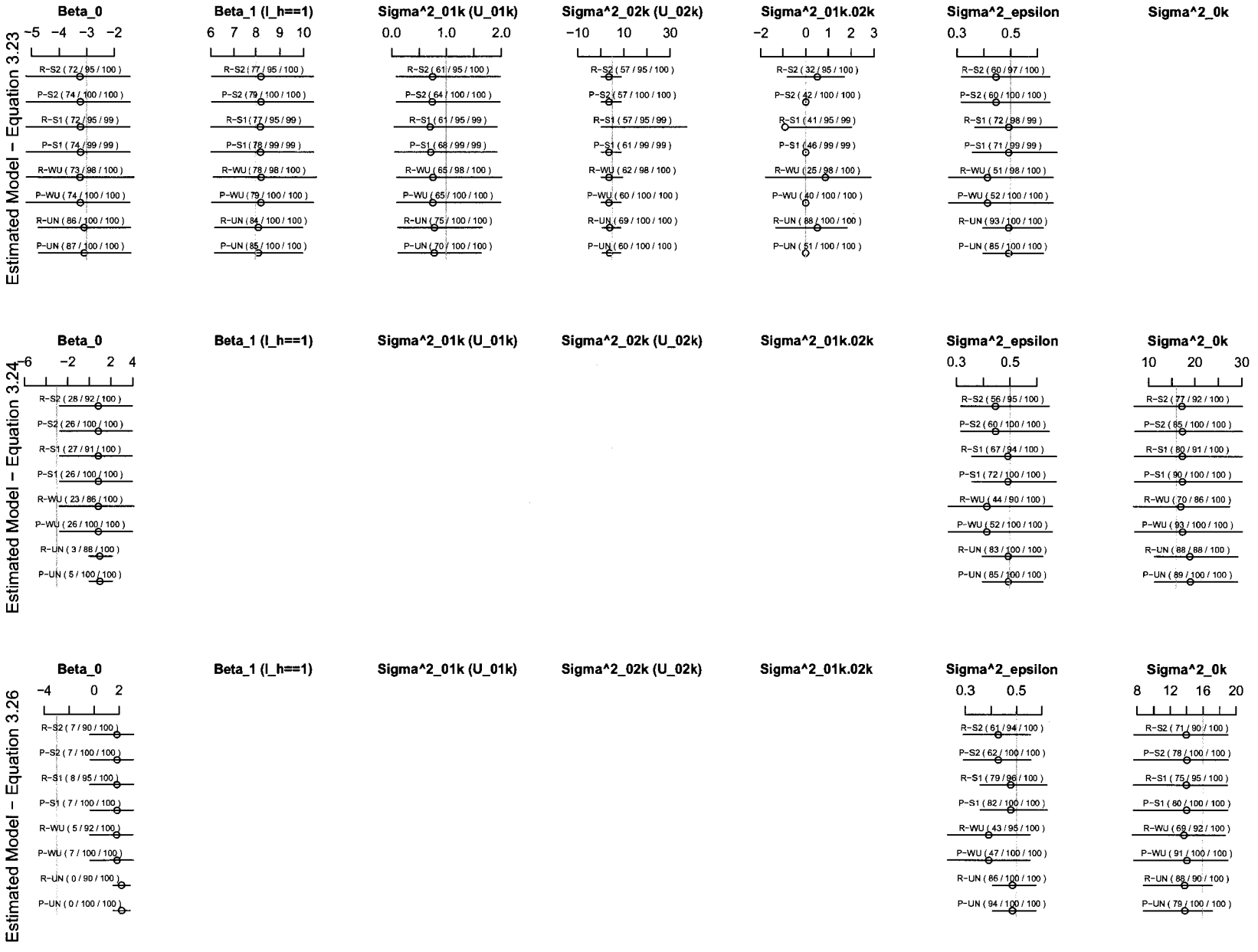

Figure 3.8: Results for Misspecification of Stratification Layers, Simulation Set 9
Generated Model - Equation 3.22

The coverage of the confidence intervals of RHS and PSHGR are similar (except for the $\sigma^2_{01k.02}$ intervals) with the RHS 95% confidence intervals for the β coefficients from the estimated model in Equation 3.23 are between 74% to 81%, and for the variance components they are between 52% and 93%. The coverage of the PSHGR 95% confidence intervals for the β coefficients from the estimated model in Equation 3.20 are between 74% to 87%, and for the variance components they are between 52% and 85%. The confidence intervals for PSHGR do not capture the $\sigma^2_{01k.02k}$ well because many of the estimated variances of the point estimates were negative. RHS was able to produce sandwich estimator variances for between 86% and 100% of the simulation runs, while PSHGR was able to produce design based estimator variances for 100% of the simulation runs.

The second and third estimated models contain model misspecification as the stratified/clustered model was reduced to a clustered model. As expected, the estimated intercept became the average of the two strata intercepts (as the sample size had 50% from each stratum) and the estimate of the random intercept includes the variance of the means of the strata and the two random effects. The estimate of the random error did not change. For more description see Section 3.7.1. The third model also includes model misspecification and informative sampling. The addition of the weights does not help compensate for the model misspecification.

The third estimated model contains informative sampling. The unweighted estimate of β_0 exhibits bias from the informative sampling. This bias is reduced by the weighted estimates, but not eliminated. We also see the bias in the unweighted estimation of σ^2_{0k}. Note that the unweighted estimate from the estimated model in Equation 3.24 is larger than

the weighted unscaled estimate from the same estimated model. However, the unweighted estimate of σ^2_{0k} from the esimtaed model in Equation 3.26 is smaller than the weighted unscaled estimate from the same simulation . We also see that all the means of the estimates of σ^2_{0k} from the estimated model in Equation 3.24 are larger than the true value, whereas for the same parameter in the estimated model from Equation 3.26 the means of the esimates are smaller than the true value. This shows again that the weighted estimates help compensate for the model misspecification, but do not eliminate it.

All the weighted estimates perform similarly for the β coefficients. The weighted estimates are all quite similar for the estimates of the variance components of the random slopes. They are closer together than the previous simulations estimates of the random error. The estimates of the variance components are exhibiting the same behavior as before, with the weighted unscaled as biased, the weighted scaled 1 overcompensating for the bias and the weighted scaled 2 between the weighted scaled 1 and the weighted unscaled estimates. The unweighted estimates a smaller 0.975, 0.025 quantile spread than the weighted estimates in all these simulations. When the spreads of the weighted estimates vary, then the weighted unscaled spread is the largest, followed by the weighted scaled 2 estimates spread and the weighted scaled 1 estimates spread.

3.4.9 Misspecification of the Stratification Layering - Clustered/Stratified Sampling - Simulation Set 10

A summary of this simulation set is in the "Mis Strat 10" column of Table 3.7. The sampling structure first samples clusters and within each cluster there are two strata. Let $I_{h==1}(I_{h==2})$ be an indicator variable that the element is in the first (second) stratum, respectively. The generating model is a random intercept model that takes into account the clustering and stratification,

$$\begin{aligned} y_{ikh} &= -3 + 8I_{h==1} + U_{0k} + \epsilon_{ikh} \qquad (3.27) \\ & U_{0k} \sim N(0,5), \quad \epsilon_{ikh} \sim N(0,0.5), \end{aligned}$$

where the effect of being in a given stratum is the same regardless of cluster membership. There are 30 population clusters, each containing two strata. Each stratum contains a random uniform number of population elements per population cluster between 25 and 50. The sample includes 5 clusters. Within each of the 5 clusters, there are two strata, and 10 elements are sampled from each stratum. Sampling of clusters is proportional to an independently generated random variable assigned to each cluster[10]. Sampling of elements within a cluster is proportional to an independently generated random variable assigned to each element[11].

There are two estimated models in this simulation set. One matches the generated

[10]Each cluster was assigned a random variable $a_k \sim$ Uniform$(-5,5)$. They were then sampled proportional to $(1+\exp(-a_k))^{-1}$.

[11]Each element was assigned a random variable $a_{ik} \sim$ Uniform$(-5,5)$. They were then sampled proportional to $(1+\exp(-a_{ik}))^{-1}$.

model, and one removes the layer of stratification to estimate a cluster only scheme,

$$y_{ikh} = \beta_0 + \beta_1 I_{h==1} + U_{0k} + \epsilon_{ikh}, \; U_{0k} \sim N(0, \sigma^2_{0k}), \; \epsilon_{ikh} \sim N(0, \sigma^2_\epsilon) \quad (3.28)$$

$$y_{ihk} = \beta_0 + U_{0k} + \epsilon_{ihk}, \; U_{0k} \sim N(0, \sigma^2_{0k}), \; \epsilon_{ihk} \sim N(0, \sigma^2_\epsilon). \quad (3.29)$$

Similar to the previous simulation set, these results are presented with the result of an additional simulation. This simulation used the same generating model, however the sampling scheme includes informative sampling for the clusters. The generating model is

$$y_{ikh} = -3 + 8 I_{h==1} + U_{0k} + \epsilon_{ikh} \quad (3.30)$$

$$U_{0k} \sim N(0, 5), \; \epsilon_{ikh} \sim N(0, 0.5)$$

and there was one estimating model,

$$y_{ihk} = \beta_0 + U_{0k} + \epsilon_{ihk} \quad (3.31)$$

$$U_{0k} \sim N(0, \sigma^2_{0k}), \; \epsilon_{ihk} \sim N(0, \sigma^2_\epsilon).$$

Sampling of clusters is proportional ot the magnitude of the random effect, U_{0k}. Sampling of elements within a cluster is proportional to an independently generated random variable assigned to each element[12]. The case in which the estimating model matched the generating model was not run due to space considerations.

The sampling scheme is missing completely at random for the estimating models in

[12]Each element was assigned a random variable $a_{ik} \sim$ Uniform$(-5, 5)$. They were then sampled proportional to $(1 + \exp(-a_{ik}))^{-1}$.

Equations 3.28 and 3.29, and it is informative for the estimating model in Equation 3.31.

Summary

The results from this simulation set are in Figure 3.9. A detailed description of the results is in Section 3.7.1.

In this simulation set, the estimation using the PSHGR method mostly matched the estimation using the RHS method. There are some differences between the PSHGR and RHS estimates that are large enough to be seen in Figure 3.9. The PSHGR weighted scaled 1 estimate of σ_ϵ^2 from the estimating model in Equation 3.29 has a much lower 0.025 quantile and mean than the corresponding RHS estimate. The PSHGR unweighted estimate of σ_{0k}^2 has a lower 0.025 quantile than the corresponding RHS estimate. The PSHGR weighted scaled 2 estimate of σ_{0k}^2 has a lower 0.975 quantile and mean than the corresponding RHS estimate. Finally, the PSHGR scaled 1 estimate of σ_ϵ^2 has a lower 0.025 quantile and mean than the corresponding RHS estimate.

The coverage of the confidence intervals of RHS and PSHGR are similar, with the RHS 95% confidence intervals of the β coefficients for the estimated model in Equation 3.28 is between 70% and 95%, and for the variance components the coverage is between 50% and 84%. The coverage of the PSHGR 95% confidence intervals of the β coefficients for the estimated model in Equation 3.28 is between 75% and 90%, and for the variance components the coverage is between 50% and 80%.

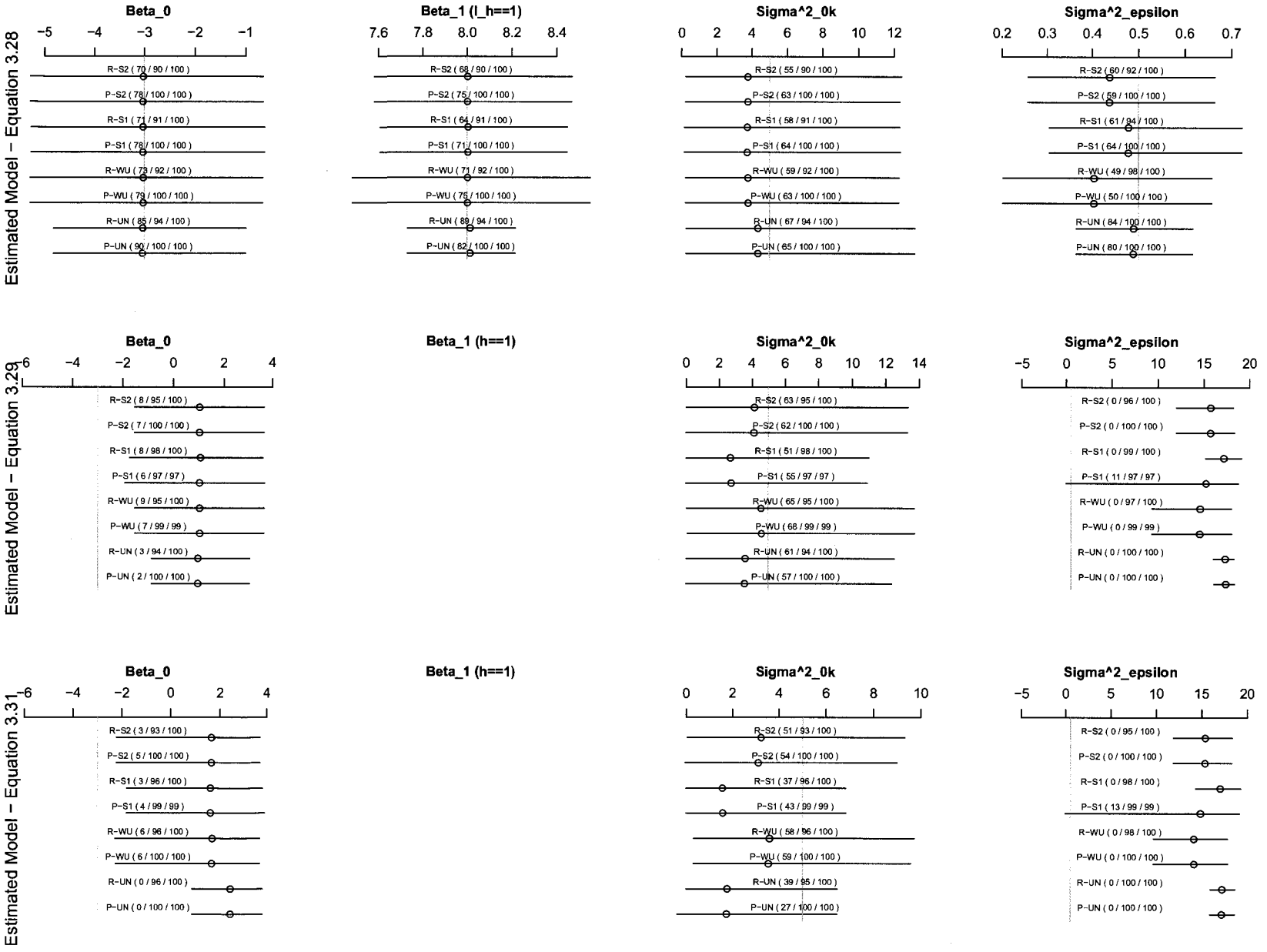

Figure 3.9: Results for Misspecification of Stratification Layers, Simulation Set 10
Generated Model - Equation 3.27

The second and third estimated models contains model misspecification. When the stratification layer was removed, it had the same effect as losing a fixed effects variable that varied according to cluster – the intercept estimate and the variance of the random error both changed. See Section 3.7.1 for more details. The addition of the weights does not help compensate for this model misspecification.

The third estimated model contained informative sampling, the effects of which can be seen in the unweighted estimates of the β_0, and σ^2_{0k} parameters. The all of the estimates (and especially the unweighted estimate) of σ^2_{0k} are smaller in the estimated model from Equation 3.31 than the corresponding estimates from the estimated model in Equation 3.29 The use of the weights helped to compensate for the informative sampling bias, but did not completely remove the bias.

All of the weighted estimates performed similarly for the β coefficients, however the variance for the weighted unscaled estimate is larger for the estimate of the stratification indicator. In addition, there are instabilities in the PSHGR estimation of σ^2_ϵ when using the scaled 1 weights. The pattern in the weighted estimates of the variance components is that the weighted scaled 1 has more bias in the same direction than the weighted unscaled estimate (except for the estimates of σ^2_ϵ when the estimated model is from Equation 3.28). The usual pattern is that the weighted scaled 1 estimates compensate (or overcompensate) for the weighted unscaled bias. Also unusual is the larger variance for the unweighted estimates of σ^2_{0k} in both the RHS and PSHGR estimates from the estimated model in Equation 3.28. The unweighted estimates a smaller 0.975, 0.025 quantile spread than the weighted estimates in all these simulations. When the spreads of the weighted estimates

vary, then the weighted unscaled spread is the largest, followed by the weighted scaled 2 estimates spread and the weighted scaled 1 estimates spread.

3.4.10 Misspecification of Stratification Layers - Stratified/Clustered/Stratified Sampling - Simulation Set 11

A summary of this simulation set is in the "Mis Strat 11" column of Table 3.7. The sampling structure is a three stage stratify/cluster/stratify scheme where each layer of stratification has two strata. Let $I_{h1==1}(I_{h1==2})$ be an indicator that the element is in the first (second) top level strata, respectively. Let $I_{h2==1}(I_{h2==2})$ be an indicator that the element is in the first (second) lower level strata, respectively. The generating model is a random intercept model that takes into account the clustering and stratification,

$$y_{ih_1kh_2} = 7 - 8I_{h_1==2} - 10I_{h_2==2} + U_{01k}I_{h_1==1} + U_{02k}I_{h_1==2} + \epsilon_{ih_1kh_2} \quad (3.32)$$

$$U_{01k} \sim N(0,1), \quad U_{02k} \sim N(0,5), \quad \epsilon_{ih_1kh_2} \sim N(0,0.5).$$

This generating model has separate means for the two top level strata where the effect of being in top level strata 1 is 5 and the effect of top level strata 2 is -3. The clusters within top level strata $h_1 == 1$ have a different random intercept variance than the clusters within top level strata $h_1 == 2$. Within each stratum / cluster, the effect of being in the bottom level second strata is the same regardless of cluster. The effect of being in lower level stratum 1 is 2 and the effect of being in lower level stratum 2 is -8. Thus the mean of a unit in the first top layer strata and the first lower level strata is 7 , the mean for the first top layer strata and the second lower level strata is -3, the mean for the second top level stratum and the first lower level stratum is -1, and the mean for the second top level stratum and the second lower level stratum is -11.

Each of the two upper level stratum contains 300 population clusters. Each cluster contains two lower level strata. Each lower level strata contains a random uniform number of population elements between 50 and 100. Within each top level strata, five clusters are sampled proportional to an independently generated random variable. Within each sampled cluster, 20 elements are sampled from each of the two strata. There are 400 elements in the sample.

There are four estimated models in this simulation set. The first estimating model matches the generating model. The second and third estimating models drop one (either the top or the bottom) layer of stratification. Finally the fourth estimating model drops both layers of stratification and has a cluster only model. The four estimated models are

$$\begin{aligned} y_{ijkl} &= \beta_0 + \beta_1 * (I_{ik} \in \text{s1}{=}\ 2) + \beta_2 * (I_{ijkl} \in \text{s2}{=}2) + U_{0k1} * (I_{ijkl} \in \text{s1}{=}1) \quad (3.33) \\ &+ U_{0k2} * (I_{ijkl} \in \text{s1}{=}2) + \epsilon_{ik} \\ & U_{0k1} \sim N(0, \sigma^2_{0k1}), \quad U_{0k2} \sim N(0, \sigma^2_{0k2}), \quad \epsilon \sim N(0, \sigma^2_\epsilon), \\ y_{ik} &= \beta_0 + \beta_1 * (I_{ijkl} \in \text{S2}{=}2) + U_{0k} + \epsilon_{ik} \quad (3.34) \\ & U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon \sim N(0, \sigma^2_\epsilon), \\ y_{ijkl} &= \beta_0 + \beta_1 * (I_{ik} \in \text{s1}{=}\ 2) + U_{0k1} * (I_{ijkl} \in \text{s1}{=}1) \\ &+ U_{0k2} * (I_{ijkl} \in \text{s1}{=}2) + \epsilon_{ik} \quad (3.35) \\ & U_{0k1} \sim N(0, \sigma^2_{0k1}), \quad U_{0k2} \sim N(0, \sigma^2_{0k2}), \quad \epsilon \sim N(0, \sigma^2_\epsilon), \\ y_{ik} &= \beta_0 + U_{0k} + \epsilon_{ik} \quad (3.36) \\ & U_{0k} \sim N(0, \sigma^2_{0k}), \quad \epsilon \sim N(0, \sigma^2_\epsilon). \end{aligned}$$

This sampling scheme is sampling completely at random for all of the estimating models.

Summary

The results from this simulation set are in Figure 3.10. A detailed description of the results is in Section 3.7.1.

In this simulation, the estimation using the PSHGR method mostly matches the estimation using the RHS method. However, there are many differences between the PSHGR and RHs estimates large enough to be seen in Figure 3.10. First consider the estimating model in Equation 3.33. The PSGHR weighted scaled 1 estimates of σ^2_{01k} have a smaller mean and a smaller 0.975 quantile than the corresponding RHS estimates. The estimates of σ^2_{02k} and $\sigma^2_{01k.02k}$ have obvious differences. For the estimating model in Equation 3.34, the PSHGR unweighted estimate of σ^2_{0k} has a larger 0.025 and 0.975 quantiles and mean than the corresponding RHS estimate. The mean of the PSHGR weighted unscaled estimates of σ^2_{0k} has a larger mean than the corresponding RHS estimates. The mean of the PSHGR weighted scaled 2 estimates of σ^2_{0k} is larger than the associated RHS estimates. There are many differences from the estimated model from Equation 3.35. The mean of the PSHGR weighted unscaled estimates of β_0 is smaller than the corresponding RHS estimate. The PSHGR weighted scaled 1 0.025 quantile for β_0 is smaller than the corresponding RHS estimate. The PSHGR mean and 0.025 quantile for the weighted unscaled estimates of β_1 are larger than the corresponding RHS estimates. The PSHGR mean and 0.975 quantiles of the weighted scaled 1 estimates of β_1 are larger than the corresponding RHS estimates.

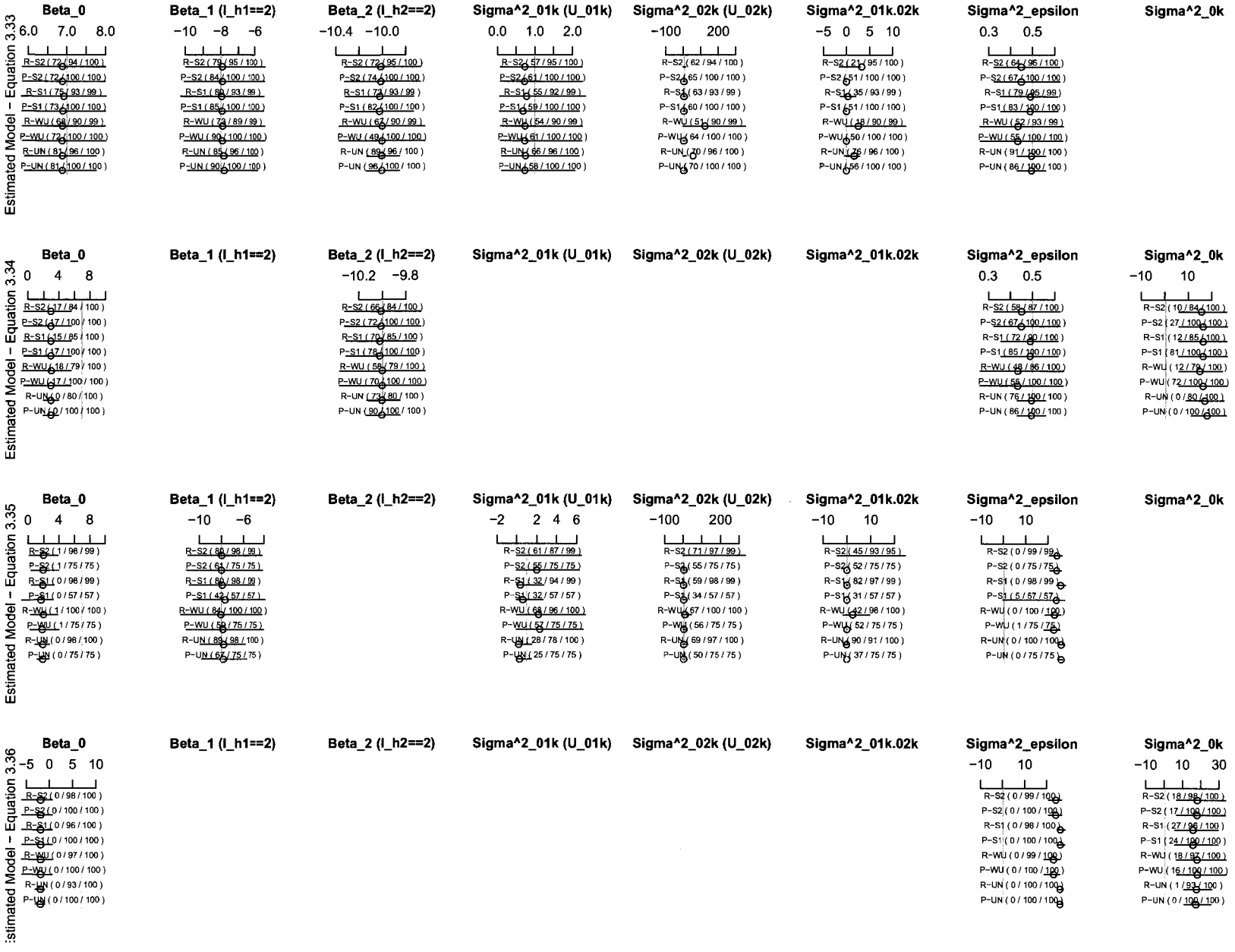

Figure 3.10: Results for Misspecification of Stratification Layers, Simulation Set 11
Generated Model - Equation 3.32

There is a large outlier in the RHS weighted scaled 2 estimates of σ^2_{01k} and σ^2_{02k} resulting in the mean of the esimates to be off of the scale of the graph. The RHS weighted unscaled estimates of $\sigma^2_{01k.02k}$have a much wider range and larger mean than the associated PSHGR estimates. Finally, the PSHGR weighted scaled 1 estimates of σ^2_ϵ have a lower 0.025 quantile and mean than the associated RHS estimates. For more details, see Section 3.7.1. The coverage of the confidence intervals of RHS and PSHGR are similar, with the RHS 95% confidence intervals for the β coefficients from the estimated model in Equation 3.33 are between 74% to 93% and for the variance components (not including $\sigma^2_{01k.02k}$) they are between 56% and 91%. The coverage of the PSHGR 95% confidence intervals for the β coefficients from the estimated model in Equation 3.33 are between 72% to 96% and for the variance components (not including $\sigma^2_{01k.02k}$) they are between 55% and 86%. RHS was able to produce sandwich estimator variances for between 75% and 100% of the simulation runs, while PSHGR was able to produce design-based estimator variances for between 57% and 100%.

The second, third and fourth rows of Figure 3.10 contain model misspecification involving dropping the top level, bottom level or both levels of stratification. The means of parameters are what is expected as described in Section 3.7.1. The misspecification is seen mostly in the estimates of β_0 and σ^2_ϵ. The addition of the weights does not compensate for this model misspecification.

There is no informative sampling in this simulation so there is no informative sampling bias.

The patterns in the different weightings are hard to see in this simulation due to the

outlying observations. For the β coefficients, it appears that the weighted unscaled estimates has a larger variance, especially for the estimating model in Equation 3.34 and 3.35. For the variance estimates, most appear to follow the pattern that the weighted unscaled estimates are biased, the weighted scaled 1 (over) compensates for the bias and the weighted scaled 2 estimates are between the weighted unscaled and the weighted scaled 1 estimates. There are two parameters where the weighted scaled 1 apears to add bias in the same direction of the weighted unscaled estimates, specifically the estimates of σ_ϵ^2 from the estimated models in Equations 3.35 and 3.36. The unweighted estimates a smaller 0.975, 0.025 quantile spread than the weighted estimates in all these simulations. When the spreads of the weighted estimates vary, then the weighted unscaled spread is the largest, followed by the weighted scaled 2 estimates spread and the weighted scaled 1 estimates spread.

3.4.11 Misspecification of Clustering Layers – Simulation Set 12

A summary of this simulation set is in the "Mis Clust 12" column of Table 3.7. The sampling structure first clusters on the top layer clusters (denoted k_1), then selects lower level clusters (denoted k_2) within the top layer clusters. The generating model is a two-level random slope model to fit the cluster/cluster design,

$$\begin{aligned} y_{ik_1k_2} &= 5 + U_{0k_1} + U_{0k_1k_2} + \epsilon_{ik_1k_2} \qquad (3.37) \\ & U_{0k_1} \sim N(0,5), \;\; U_{0k_1k_2} \sim N(0,1), \;\; \epsilon_{ik_1k_2} \sim N(0,0.5). \end{aligned}$$

There are 30 top level population clusters and within each top level population cluster there are 10 bottom level population clusters with a random uniform number of population units per cluster between 25 and 50. The sample contains 5 top level clusters, 5 bottom level clusters and 3 elements per bottom level cluster. The top level clusters are sampled proportional to first independent random variable, the bottom level clusters are sampled proportional to a second independent random variable, and the elements within the bottom cluster are sampled proportional to a third independently generated random variable. There are two estimating models in this simulations set, the first removes the bottom layer of clustering,

$$\begin{aligned} y_{ik_1k_2} &= 5 + U_{0k_1} + \epsilon_{ik_1k_2} \qquad (3.38) \\ & U_{0k_1} \sim N(0, \sigma^2_{0k1}), \;\; \epsilon_{ik_1k_2} \sim N(0, \sigma^2_{\epsilon}), \end{aligned}$$

and the second removes the top layer of clustering,

$$y_{ik_1k_2} = 5 + U_{0k_1k_2} + \epsilon_{ik_1k_2} \tag{3.39}$$
$$U_{0k_1k_2} \sim N(0, \sigma^2_{0k1k2}), \quad \epsilon_{ik_1k_2} \sim N(0, \sigma^2_\epsilon).$$

Due to time constraints, none of the estimated models match the generating model. This sampling scheme is sampling completely at random for both estimated models.

Summary

The results from this simulation set are in Figure 3.11. For a complete description of the simulation results, see Section 3.7.1.

In this simulation, the estimation using the PSHGR method mostly matched the estimation using the RHS method. THere are some differences between the PSHGR and RHS estimates large enough to be seen in Figure 3.11. From the estimated model in Equation 3.38, the PSHGR empirical confidence intervals for the weighted scaled 1 and weighted scaled 2 estimates of σ^2_{0k1} are longer than the corresponding RHS intervals. The 0.025 quantile of the PSHGR weighted scaled 1 and weighted scaled 2 estimates of σ^2_ϵ are smaller than the corresponding RHS quantiles. For the estimated model in Equation 3.39, the 0.975 quantiles for the weighted unscaled, weighted scaled 1 and weighted scaled 2 are larger for the RHS intervals than for the corresponding PSHGR intervals.

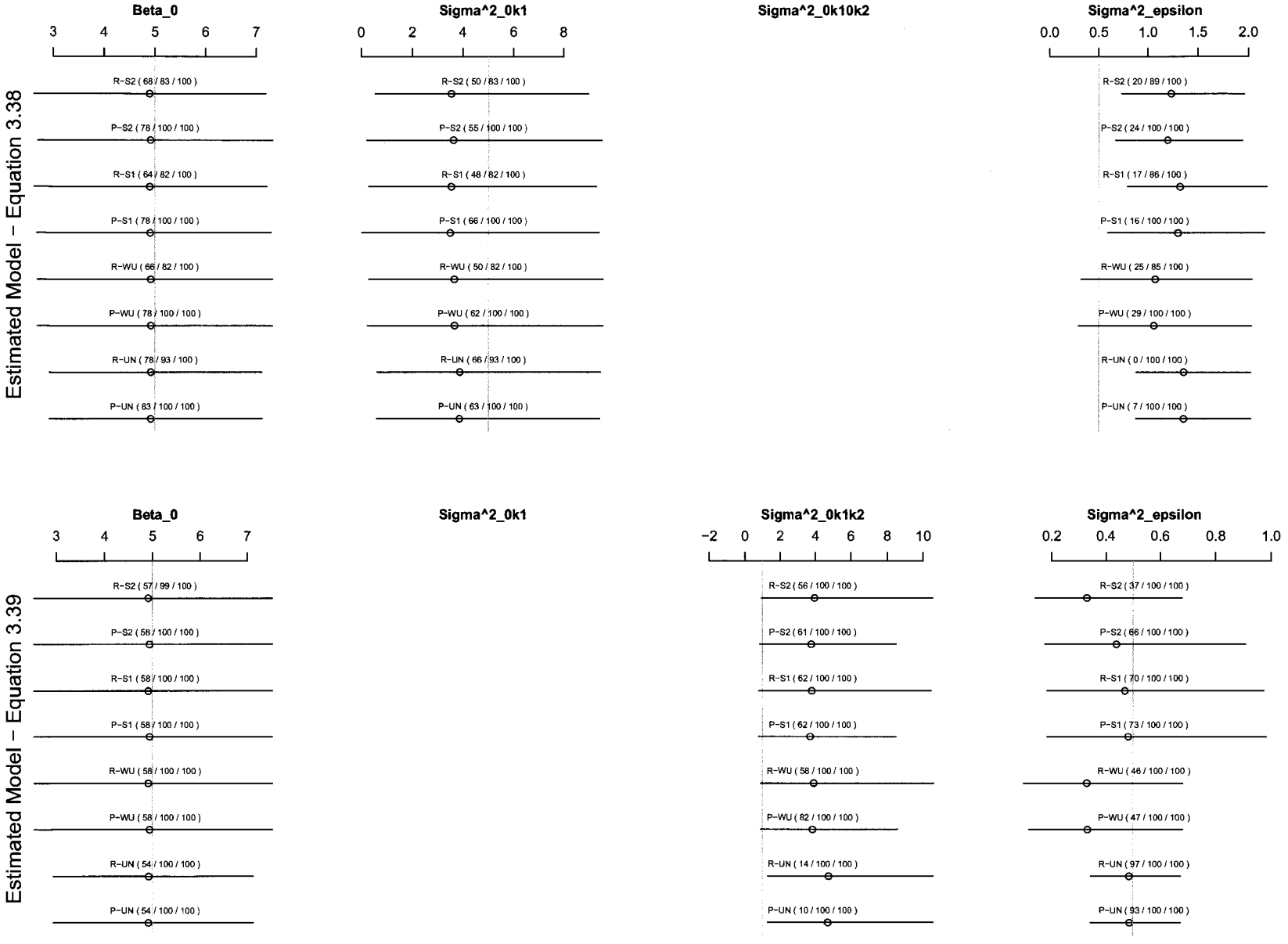

Figure 3.11: Results for Misspecification of Clustering Layers, Simulation Set 12
Generated Model - Equation 3.37

For the PSHGR 0.025 quantile of the weighted unscaled estimate of σ_ϵ^2 is larger than the corresponding RHS quantile. The mean of the PSHGR scaled 2 estimates of σ_ϵ^2 is larger than the corresponding RHS mean. Finally, the 0.025 and 0.975 quantiles and the mean of the PSHGR weighted scaled 2 estimates of σ_ϵ^2 are larger than the corresponding RHS estimates. For more details, see Section 3.39.

The coverage of the confidence intervals of RHS and PSHGR is not analyzed in this simulation as both estimated equations contain model misspecification.

The first and second rows both contain model misspecification, as the generating model is a three level random intercept model and the two estimating models are two level random intercept models. In these simulations, the variance from the cluster level that was dropped was merged into the remaining cluster level or the random error term. For a description of the expected results, see Section 3.7.1. The addition of the weights did not compensate for the model misspecification.

There is no informative sampling in this simulation so there is no informative sampling bias.

The means of all weighted estimates are similar for the β coefficient. The variance of the unweighted estimates is smaller than for the weighted estimates. For the σ_ϵ^2 from the estimated model in Equation 3.39, we see the pattern where the weighted unscaled estimates are biased, the scaled 1 estimates (over) compensate for the bias and the scaled 2 estimates are between the weighted unscaled and the weighted scaled 1 estimates. However, for the estimates of the variance components from the estimated model in Equation 3.38, we see that the scaled 1 weights are adding bias in the same direction as the weighted

unscaled weights. With the differences in estimates of $\sigma^2_{0k2.0k2}$, the pattern is difficult to determine. The unweighted estimates a smaller 0.975, 0.025 quantile spread than the weighted estimates in all these simulations. When the spreads of the weighted estimates vary, then the weighted unscaled spread is the largest, followed by the weighted scaled 2 estimates spread and the weighted scaled 1 estimates spread.

3.5 Mean Squared Error Comparisons of the Simulations

It is not clear how to compare the different methodologies (PSHGR vs. RHS) crossed by the different weightings. A criterion such as AIC or BIC is desired, however it is not clear if these are appropriate. AIC and BIC aid in model selection, however the insertion of the sampling weights in different places doesn't necessarily fall into model selection. To help find a good metric, I propose two different calculations based on the mean squared error. I evaluate the simulations based on their metrics and discuss the strengths and weaknesses. I do not believe these are good metrics to evaluate the simulations, but they identify issues that need to be considered when determining a metric.

Relative Square Root Mean Squared Error (RRMSE)

Let $\hat{\beta}_1$ be the estimate of β_1 and let n be the number of simulation runs that produced point estimates. Then $RRMSE = \sqrt{\sum_{i=1}^{n} n^{-1}\beta_1^{-2}(\hat{\beta}_1 - \beta_1)^2}$. This is the square root of the mean squared error that is scaled by the magnitude of the parameter. This metric balances the bias and the variance for each parameter.

RRMSE is a measure of the model misspecification and informative sampling. Often, the model misspecification dominates the $RRMSE$. To help compensate for this, the $ARRMSE$ is also computed.

Adjusted Relative Square Root Mean Squared Error (ARRMSE)

Similar to the $RRMSE$, however instead of using the true value of the parameter we use the anticipated value of the parameter value given the model misspecification. For example, if β_{1A} is the anticipated value of the parameter given the model mis-

specification, then $ARRMSE = \sqrt{\sum_{i=1}^{n} n^{-1}\beta_{1A}^{-2}(\hat{\beta}_1 - \beta_{1A})^2}$ where n is the number of simulation runs out of 100 that produced point estimates. This $ARRMSE$ removes the model misspecification component from the $RRMSE$, and measures the effects of informative sampling.

For more information on the derivation of the anticipated parameter values, see the description for the simulation in question in Section 3.7.1. The anticipated parameter values are tabulated in Section 3.7.5.

The values of the $RRMSE$ and $ARRMSE$ for each estimate in the simulations are in Section 3.7.5. To summarize this data, I added the $RRMSE$ ($ARRMSE$) values of each estimate for a given estimating model, methodology (PSHGR vs. RHS) and weighting scheme. This has advantages and disadvantages. The advantage is that when a model is estimated, the estimates of the parameters that are used must come from one estimating set. For example, I can not choose an estimating model and the estimate the fixed effects using, for example, PSHGR unweighted estimates and then estimate the random effects using RHS weighted scaled 2 estimates. This merges all estimates from a given estimated model together within one framework. The problem is that when the scales differ and when there is model misspecification, the estimate of one parameter in the model can dominate the mean squared error calculation. For that reason, the relative MSE is used (i.e. diving by the true/anticipated value) and both $RRMSE$ and $ARRMSE$ are presented. However, when the true (or anticipated value) is zero, then the $RRMSE$ (or $ARMSE$) can not be computed.

Table 3.8 contains a summary of the results of the 12 simulation sets. The first column contains the name of the simulation set. The second column contains the equation number of the estimated equation. The $RRMSE$ for all the parameters of a given type of weight and method are then added together. In the subsequent columns, P and R in the given column represent the PSHGR and RHS weighting scheme that produces the smallest $RRMSE$, and PA and RA represent the smallest $ARRMSE$. Note that for estimated models in Equations 3.12, 3.15, 3.18 and 3.21 a random intercept is included in the estimated model instead of the random slope. Because the true parameter value of the random intercept variance is zero, the $RRMSE$ can not be computed however the $ARRMSE$ is computed and recorded. For the estimated models in Equations 3.23, 3.33 and 3.35, the true parameter value of $\sigma^2_{01k.02k}$ is zero. For these equations, the $RRMSE$ and $ARRMSE$ are computed without a contribution from the estimates of $\sigma^2_{01k.02k}$. More detailed summary tables are in Section 3.7.4

For a detailed description of the MSE results, see Section 3.7.4. The same weighting method generally produced the lowest MSE for both the PSHGR and RHS methodologies. When this is not the case (see table 3.8 for Equation numbers 3.4 and 3.31) it is due to differences in the methods described in Section 3.7.1.

The unweighted estimates generally provided the lowest $ARRMSE$. The cases where this is not true (see table 3.8 for Equation numbers 3.8 and 3.31) are due to informative sampling. The $AARMSE$ prefers the unweighted estimates due to their smaller variance. The bias induced by the informative sampling in these simulations is not large enough to penalize the unweighted estimates. The $RRMSE$ is more sensitive to model misspecifica-

		Eqn. Num.	Weighting Scheme with Lowest MSE: Unweighted	Weighted Unscaled	Weighted Scaled 1	Weighted Scaled 2
Mis Fix 1		3.3		P R		
		3.4	PA RA	R	P	
		3.5	PA RA	P R		
Mis Fix 4		3.7	P R			
		3.8	P R		PA RA	
		3.9	P R PA RA			
Mis Ran 5		3.11	P R			
		3.12	PA RA			
Mis Ran 6		3.14				P R
		3.15	PA RA			
Mis Ran 7		3.17	P R			
		3.18	PA RA			
Mis Ran 8		3.20			P R	
		3.21	PA RA			
Mis Strat 9		3.23	P R			
		3.24	PA RA			
		3.26	PA RA			
Mis Strat 10		3.28	P R			
		3.29	PA RA			
		3.31			RA	PA
Mis Strat 11		3.33	P		R	
		3.34	PA RA			
		3.35	PA RA	P R		
		3.36	PA RA			
Mis Clust 12		3.38	PA RA	P R		
		3.39		PA RA P R		

Table 3.8: Mean Squared Errors for each Simulation Set

tion and appears to prefer the unweighted and weighted unscaled estimates. The preference for the weighted unscaled estimates occurs because these estimates show the most bias in the variance components. When the model is misspecified and the anticipated value of the variance component gets large causing a very large bias when compared to the true value of the parameter. That variance component dominates the sum of the $RRMSE$ and it is often the weighted unscaled estimates that are closest to the true value of the parameter. Ffor example, see the estimate of σ_ϵ^2 in Figure 3.2 from the estimated model in Equation 3.5.

As described in Section 3.7.4, the level of informativeness is a big factor as to which type of weighting scheme is preferred.

3.6 Simulation Result Summary

This chapter provides new contributions or supports existing claims on each of five goals. This is accomplished through 12 sets of simulations that compare estimation methods (PSHGR or RHS) and scaling of weights (unweighted, weighted unscaled, weighted scaled 1 and weighted scaled 2) on correctly and incorrectly specified models, both with and without informative sampling.

The first goal is to compare the results from the different methods of inserting weights into LME models. This chapter compares the method of Rabe-Hesketh and Skrondal (2006) , which is the same as Asparouhov (2006), to the method of Pfeffermann et al. (1998). These simulations found that the RHS and PSHGR methods provide remarkably similar results. When the results are not similar, it is mostly due to sensitivities of the numerical quadrature to the number of quadrature points in the `gllamm()` function that implements the RHS method. Neither RHS nor PSHGR provided this direct comparison in their papers.

The second goal is to compare the sandwich estimator (used by RHS) and the design-based estimator (used by PSHGR) when obtaining the variances of the point estimates. When there is no model misspecification, the confidence intervals based on the sandwich estimator have similar coverage levels as the confidence intervals based on the design-based estimates. However, when there is model misspecification, the design-based confidence intervals have coverage that is unexpectedly large, implying that the variance estimates are too large. Neither RHS nor PSHGR provided a comparison in their papers and neither of them looks at the performance of the variance estimators in the presence of model

misspecification.

The third goal of this chapter is to investigate the assertion that adding sampling weights can compensate for model misspecification in LME models. The simulations in this chapter indicate that the weights can help for model misspecification only when the model misspecification induces informative sampling. Bias related to a misspecified model that does not relate to the sampling design is unaffected by the sampling weights. Previous simulation studies did not study model misspecification.

The fourth goal of this chapter is to investigate the assertion that adding sampling weights can compensate for informative sampling in LME models. The simulations in this chapter support those conclusions. The inverse probability sampling weights can help compensate for bias induced by informative sampling, though they do not eliminate the bias. This supports the conclusions in the previous simulation papers.

The final goal of this chapter is to investigate the different scalings of the weights, which are introduced in Section 2.4.2. These simulations found that the unweighted estimates have the smallest variance. However, when there is informative sampling, the unweighted estimates are biased. The weighted unscaled estimate corrects the bias in the fixed effects, but produces more bias in the random effects. The weighted scaled 1 estimates remove the bias in the fixed effects, and correct (or overcorrect) for the weighted unscaled bias in the random effects. The weighted scaled 2 estimates remove the bias in the fixed effects and have a bias between the weighted unscaled and weighted scaled 1 estimates in the random effects. There are times when the scaled 1 estimates have more bias in the same direction as the weighted unscaled estimates. The conditions upon which this occurs need

to be further investigated. RHS, PSHGR and ASP tentatively recommended the weighted scaled 2 estimates. These simulations provide a good characterization of the relationship between the scaled estimates and demonstrate that the variance of the scaled 1 estimates is sometimes lower than the scaled 2 estimates.

Comparison of the weighting schemes over the different estimated models is difficult. To gain insight into the comparison, I computed the $RRMSE$ and $ARRMSE$ metrics and looked at their strengths and weaknesses. The $RRMSE$ metric incorporates informative sampling, model misspecification and variance and generally prefers the unweighted or weighted unscaled estimates. This is due to the low variance of the unweighted estimates and the pattern of bias in the weighted unscaled estimates. The $ARRMSE$ metric incorporates informative sampling and variance, and generally prefers the unweighted or weighted scaled 1 estimates. This is due to the low vairance of the unweighted estimates, and the slightly higher variance, but lower bias of the weighted scaled 1 estimates. None of the previous simulation papers attempted a metric across all estimates in a model.

This chapter contributes a new way to view the simulation results. RHS, ASP, KG and PSHGR produced tables of numbers that are difficult to read and make quick comparisons. The stacked line interval format of the displays in this chapter provides a quick visual way to compare all methods together and across multiple simulations.

The results of this chapter can be generalized to more complex LME models. This chapter addressed the effects of model misspecification and informative sampling on fixed effects in two scenarios; 1) biases confined to one one level (by removal of either the x_{1k} or x_{2ik} fixed variables in simulation sets 1 and 4, for example) and 2) biases spread across

levels (by the removal of the random slope on x_{2ik} in simulation set s 7 and 8, for example). These scenarios can be easily generalized into more complex models. As the random effect structure increases and becomes more complex, I speculate that the bias in the random effects will become worse. This is because the ML estimates of the random effects are biased where the bias of one variance component depends on other variance components, as seen in the case of the estimates of σ^2_{0k} in simulation set 4 with the estimated model in Equation 3.9. The use of the weights adds to the bias of the random effects. Once one random effect estimate is biased, that bias may be propagated through to other variance components.

3.7 Appendix

3.7.1 Description of Simulation Results

Result Description for Misspecification of Fixed Variables - Simulation Set 1

We want to flag if there are large differences between the PSHGR and RHS estimates for a given iteration. To do this, the standard deviation of the parameter estimate over the 100 iterations is obtained separately for the PSHGR and the RHS estimates. The smaller of these standard deviations is used as a threshold to flag "large" differences between PSHGR and RHS estimates. For each iteration, the difference between the PSHGR and the RHS estimates is compared to the threshold to identify estimates where the difference is greater than one standard deviation. Unless otherwise mentioned, the difference between the PSHGR and RHS estimates is less than the threshold.

Figure 3.12 contains a plot of the weighted scaled 1 estimates for β_2, the unweighted and weighted unscaled estimates of σ^2_{0k} and the unweighted estimates of σ^2_ϵ from the estimated model in Equation 3.4. The solid black lines are the upper and lower thresholds. From the figure, we see that there is one point that is outside the lines for the estimate of β_2, 11 and 6 points outside the lines for the unweighted and weighted unscaled estimates of σ^2_{0k} respectively, and one point outside the line for the estimate of σ^2_ϵ. The differences between the weighted scaled 1 estimates for β_2 are too small to be seen in Figure 3.2, however the differences in unweighted and weighted unscaled estimates for σ^2_{0k} can be seen as the means do not match each other. The differences in the unweighted estimates of σ^2_ϵ can also be seen in Figure 3.2. It appears that the means may be different for the weighted

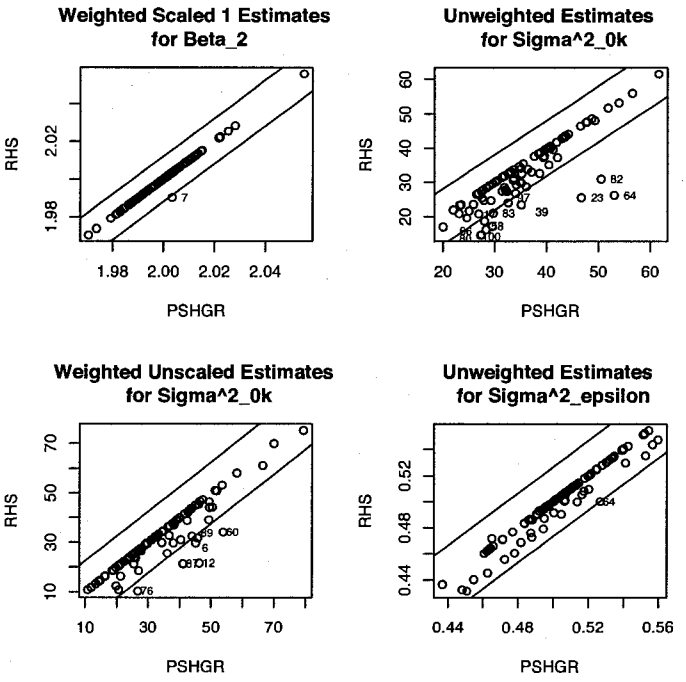

Figure 3.12: Comparison of PSHGR vs. RHS for Estimates from Equation 3.4

scaled 2 estimate of σ^2_{0k}, however all the individual estimation differences are less than the threshold.

For the estimates from the estimating model in Equation 3.5, note that the PSHGR and RHS estimates using the scaled 1 weights do not have 100 estimates. For PSHGR, simulation runs 21 and 94 did not converge in 500 iterations. For RHS, simulation runs 28 and 41 did not converge when the number of quadrature points were increased from 15 until 30. Figure 3.13 contains the unweighted estimates of σ^2_{0k} and the weighted scaled 1 estimate of σ^2_ϵ. For the estimates of σ^2_{0k}, there are a number of PSHGR estimates that range from 0 to 2 while the RHS estimates are all about 0.25. I believe that this is a

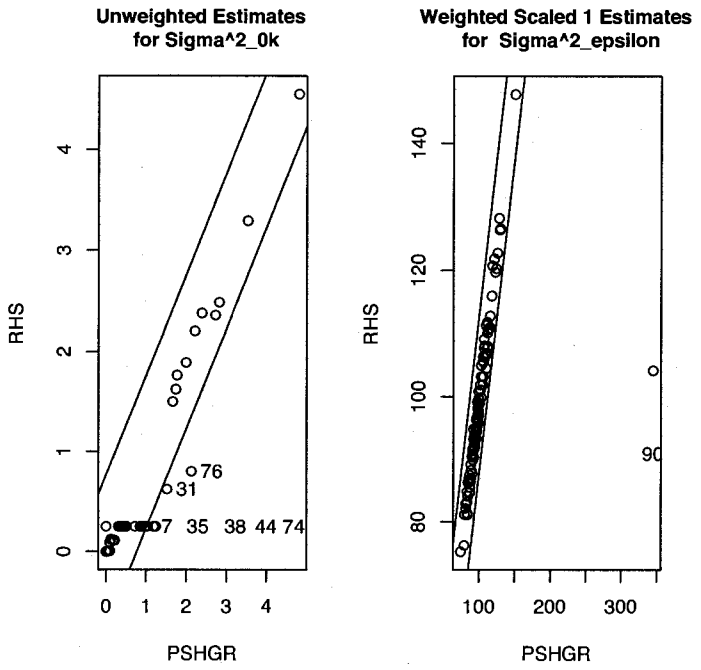

Figure 3.13: Comparison of PSHGR vs. RHS for Estimates from Equation 3.5

problem with the RHS estimation, however his pattern should be looked into further. For the estimate of σ_ϵ^2 the PSHGR weighted scaled 1 estimate of σ_ϵ^2 for simulation run 90 is 345. I believe this is an instability with the PSHGR estimation and should be looked into further. The differences in the PSHGR and RHS estimates of σ_{0k}^2 can not be seen in Figure 3.2, however the estimates of σ_ϵ^2 appear to have different means in the figure.

We next determine what we would expect the results to be for each of the estimating models. The top row of Figure 3.2 contains the summary of the estimated model from Equation 3.3, where the estimated model matches the generating model. We know that the unweighted estimates of β should be unbiased based on Section 6.2 of Searle et al.

(1992). All of the estimation methods have minimal bias and comparable quantiles for the β parameters. We also know from Section 2.4.2 that the variance components are not necessarily unbiased. Specifically, the σ^2_{0k} parameter depends on the intra-class correlation. The intra-class correlation in this data set was $\frac{0.2}{0.2+0.5} = 0.29$, and the simulation results show a slight positive bias for the weighted estimate using unscaled weights. For weighted estimates using scaled 1 and scaled 2 weights, the biases are negative and approximately the same magnitude. The σ^2_ϵ parameter follows the trends outlined in Section 2.4.2, with the weighted unscaled estimates having larger negative bias, the weighted scaled 1 and weighted scaled 2 estimates having smaller positive bias.

In the middle panel, the estimated model from Equation 3.4 no longer contains the x_{1k} variable. The mean of the missing $-2x_{1k}$ term is -6, resulting in a new intercept estimate of 1-6=-5. In addition, the variance of the missing $-2x_{1k}$ term is 36, resulting in a new σ^2_{0k} estimate of approximately 36+0.5=36.5. The bias in σ^2_{0k} follows the same trend as first row of the simulation results. The bias in σ^2_ϵ is larger than from the model in Equation 3.3. As expected, the weighting does not do anything to help in this model misspecification.

In the bottom panel, the estimated model from Equation 3.5 no longer contains the x_{ik} variable. The mean of the missing $2x_{2ik}$ term is 2, resulting in a new intercept estimate of approximately 3. The variance of the missing $2x_{2ik}$ term is 100, resulting in a new σ^2_ϵ estimate of approximately 100.5. The estimate of σ^2_{0k} is more difficult to predict, as the unweighted and weighted scaled 1 estimates of σ^2_{0k} are occasionally negative. For computations on how to predict these values, see Section 3.7.2 below. For the unweighted estimates (based on the calculations in Section 3.7.2), $E(\hat{\sigma}^2_{0k}|\hat{\sigma}^2_{0k} \geq 0)$ is computed as the

average of the 39 non-negative estimates, which is 0.79. We can compute $p = \Pr(\hat{\sigma}^2_{0k} < 0) \approx 0.61$ by assuming that this is a balanced simulation with the number of clusters as 35, the number of elements per cluster as 20, $\sigma^2_{0k} = 0.2$ and $\sigma^2_\epsilon = 100.5$. As a result $E(\hat{\sigma}^2_{0k}) = (1-p)E(\hat{\sigma}^2_{0k}|\hat{\sigma^2_{0k}} \geq 0) \approx 0.39 * 0.79 = 0.31$. Thus our theoretic estimate σ^2_{0k} is 0.31. Compare this to our actual results by allowing all of the negative $\hat{\sigma}^2_{0k} = 0$. We get an estimate of σ^2_{0k} over the 100 iterations of 0.31. Thus the simulated result for $\hat{\sigma}^2_{0k}$ matches the theoretical result. The scaled 1 case is computed similarly, but with 67 of the 100 iterations producing negative estimates of σ^2_{0k}. $E(\hat{\sigma}^2_{0k}|\hat{\sigma^2_{0k}} \geq 0)$ is computed as the average of the 33 non-negative estimates, which is 1.63. We can compute $p = \Pr(\hat{\sigma^2_{0k}} < 0) \approx 0.67$ by assuming that this is a balanced simulation with the number of clusters as 35, the number of elements per cluster as 20, $\sigma^2_{0k} = 0.2$ and $\sigma^2_\epsilon = 100.5$. As a result $E(\hat{\sigma}^2_{0k}) = (1-p)E(\hat{\sigma}^2_{0k}|\hat{\sigma^2_{0k}} \geq 0) \approx 0.33 * 1.63 = 0.54$. Thus our theoretical estimate of σ^2_{0k} is 0.54. Compare this to our actual results by allowing all of the negative $\hat{\sigma}^2_{0k} = 0$. We get an estimate of σ^2_{0k} over the 100 iterations of 0.52. It is assumed that the difference between 0.54 and 0.52 is due to the fact that this is not a balanced simulation.

Following Section 2.4.2, the approximate upper bound of the bias in the estimate of σ^2_{0k} in the scaled 2 case is $-\frac{\sigma^2_{0k}}{K} + \sigma^2_\epsilon(\frac{n_1^2}{N_1^2} - \frac{1}{N_1}) = 5.8$. The mean scaled 2 estimate for RHS is 7.6, which has a bias of 7.1. There is an unexplained bias of 7.1-5.8=1.3, which I assume is attributed to the non-balanced nature of this simulation. In the weighted unscaled case, assume that the number of population elements per cluster is 75 (it is between 50 and 100), the number of sampled elements is 20 and $\sigma^2_\epsilon = 100.5$. Then the bias bounds are approximately -5 to 94. The bias from the simulations is approximately 10, so the bias is within

what is expected. Note that the ICC is now fairly small (0.2/100.7=0.002). As expected, the weighted estimates did not appear to compensate for the model misspecification.

Result Description for Misspecification of Fixed Variables – Simulation Set 4

We want to flag if there are large differences between the PSHGR and RHS estimates for a given iteration. To do this, the standard deviation of the parameter estimate over the 100 iterations is obtained separately for the PSHGR and the RHS estimates. The smaller of these standard deviations is used as a threshold to flag "large" differences between PSHGR and RHS estimates. For each iteration, the difference between the PSHGR and the RHS estimates is compared to the threshold to identify estimates where the difference is greater than one standard deviation. Unless otherwise mentioned, the difference between the PSHGR and RHS estimates is less than the threshold. In this set of simulations, there are a number of datasets that were problematic for all weighting schemes and all parameters. For example, when the estimating model is in Equation 3.7, the difference between the PSHGR and RHS estimates is greater than the threshold for all estimates for the data from simulation run 18. The plots to show these differences for each parameter and each scaling are not shown to conserve space.

When the estimating model is from Equation 3.8, the only parameter that produces differences between the PSHGR and RHS estimates that are greater than one threshold is σ^2_{0k}, as shown in Figure 3.14. For this parameter, for the unweighted estimates simulation run 80 is larger than the threshold, for the weighted unscaled estimates simulation runs 15 and 79 are larger than the threshold and for the weighted scaled 2 estimates simulation runs 36 and 63 are larger than the threshold.

When the estimating model is from Equation 3.9, the simulation runs 1, 84 and 96 produced differences between PSHGR and RHS larger than the threshold in many the

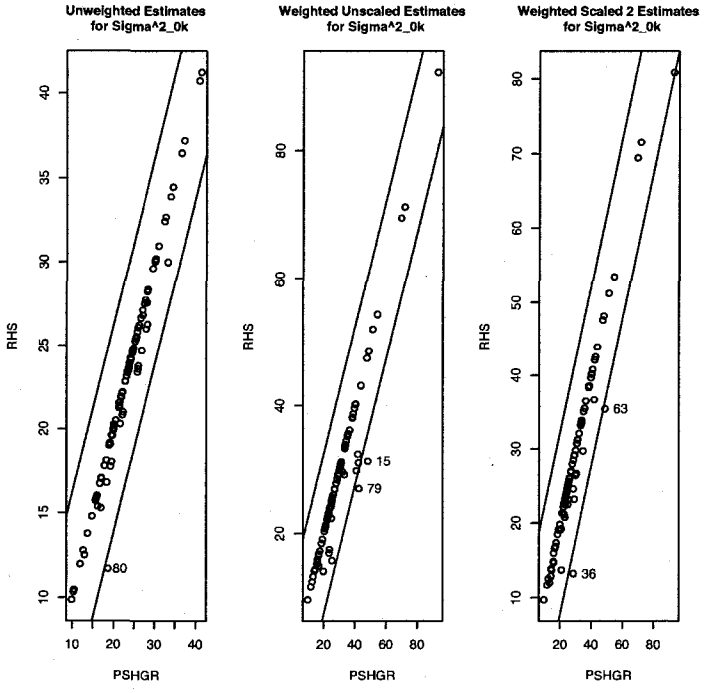

Figure 3.14: Comparison of PSHGR vs. RHS for Estimates from Equation 3.8

parameter estimates. The plots to show these differences for each parameter and each scaling are not shown to conserve space. However, there were some notable differences in PSHGR and RHS in the unweighted and weighted scaled 1 estimates of σ^2_{0k}, as seen in Figure 3.15. Similar to what was seen in Figure 3.13, the PSHGR estimates appear to vary between 0 and 1 (or 0 and 2) while the RHS estimates are 0.25. I believe that this is a problem with the RHS estimation, however his pattern should be looked into further.

We next determine what we would expect the results to be for each of the estimating models. The top row of Figure 3.3 contains the summary of the estimating model from Equation 3.7. When the estimated model matches the generating model, all of the esti-

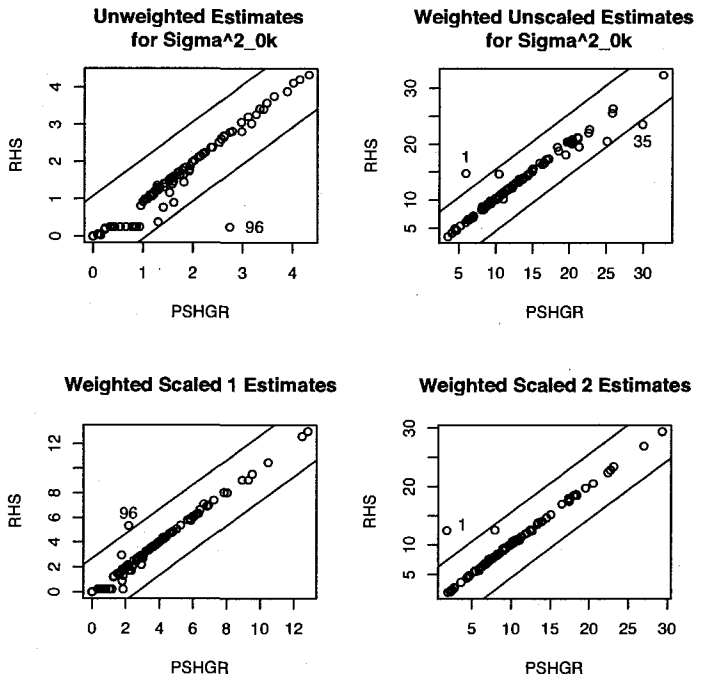

Figure 3.15: Comparison of PSHGR vs. RHS for Estimates from Equation 3.9

mation methods (PSHGR, RHS for all of unweighted, weighted unscaled, weighted scaled 1 and weighted scaled 2) have minimal bias for the β parameters. The weighted estimates have larger spreads than the unweighted estimates. In addition, the weighted unscaled estimates appear to have a larger variance than the other weighted methods for the estimation of β_2. The σ^2_{0k} parameter follows the trends outlined in Section 2.4.2. The simulation results show minimal bias for the unweighted and weighted scaled 2 estimates, a slight positive bias for the weighted unscaled estimate and a negative bias for the weighted scaled 1 estimates. It appears that the σ^2_ϵ parameter follows the trends outlined in Section 2.4.2, with the weighted unscaled estimates having larger negative bias, the weighted scaled 1

estimates have minimal bias and the weighted scaled 2 estimates are in between them. The unweighted estimates are also unbiased.

The middle panel of Figure 3.3 contains the summary of the estimating model from Equation 3.8. The estimated model no longer contains x_{1k}. This is a case of informative sampling as the clusters are sampled according to the size of x_{1k}. The mean of the missing $-2x_{1k}$ term would be -6 if there were not informative sampling, which would change the estimate of the intercept to be approximately 1-6=-5. However, because larger x_{1k} are oversampled, the expected value of $-2x_{1k}$ is more negative in the sample than in the population. This is reflected in the unweighted estimates with an average intercept of approximately -8. The weighted estimates help to compensate for the informative sampling, as they all have an intercept estimate of approximately -5. The estimates of β_2 are unaffected by the model misspecification and informative sampling. The variance of the missing $-2x_{1k}$ term would be 36 if there were no informative sampling. With no informative sampling, we would expect the estimate of σ^2_{0k} to be approximately 0.2+36=36.2. However, because the larger x_{1k} are oversampled, the variance of x_{1k} in the sample is less than the variance of x_{1k} in the population. This is reflected in the estimation of σ^2_{0k} because the unweighted estimates are smaller than the weighted estimates. Note that the mean of the weighted estimates is approximately 29, which is still smaller than the mean of the weighted estimates from the estimated model in Equation 3.5 without informative sampling, which was approximately 33. The estimates of σ^2_ϵ are not affected by this model misspecification.

The bottom panel of Figure 3.3 contains the summary of the estimating model from Equation 3.9. The estimated model no longer contains x_{2ik}. This is a case of informative

sampling as the units are sampled according to the size of x_{2ik}. The mean of the missing $2x_{2ik}$ term would be 2 if there were not informative sampling, that would change the estimate of the intercept to be approximately 1+2=3. However, because larger x_{2ik} are oversampled, the expected value of $2x_{2ik}$ is larger in the sample than in the population. This is reflected in the unweighted estimates with an average intercept of approximately 9. The addition of the weights helps to compensate for the informative sampling, with intercept estimates of between 3.5 and 4.0. Note that these are still larger than the estimates from the estimation of the intercept from Equation 3.5 where there was no informative sampling. The estimation of β_1 is unaffected by the informative sampling and model misspecification. The variance of the missing $2x_{2ik}$ term would be 100 if there were no informative sampling. This variance is added to the estimate of σ^2_ϵ for an estimate of about 100+0.5 = 100.5 when there is no informative sampling. However, because the larger x_{2ik} are oversampled, the variance of x_{2ik} in the sample is less than the variance of x_{2ik} in the population. This smaller variance is reflected in the unweighted estimates (especially when compared to the the results from the estimated model in Equation 3.5 where the unweighted estimates are larger than the weighted unscaled estimates). The estimates of σ^2_{0k} are larger than when all covariates are in the model and this is due to the smaller intra-class correlation, similar to the situation from the estimated model in Equation 3.5. However, the estimates of σ^2_{0k} in this simulation set are larger than the estimates from the estimated model in Equation 3.5.

Results Description for Misspecification of Random Variables – Simulation Set 5

We want to flag if there are large differences between the PSHGR and RHS estimates for a given iteration. To do this, the standard deviation of the parameter estimate over the 100 iterations is obtained separately for the PSHGR and the RHS estimates. The smaller of these standard deviations is used as a threshold to flag "large" differences between PHSGR and RHS estimates. For each iteration, the difference between the PSHGR and the RHS estimates is compared to the threshold to identify estimates where the difference is greater than one standard deviation. Unless otherwise mentioned, the difference between the PSHGR and RHS estimates is less than the threshold. There were no differences larger than the threshold for the estimated model in Equation 3.11. Figure 3.7.1 contains the simulation runs in which the estimates of PSHGR and RHS are larger than the threshold for the estimated model in Equation 3.12. These occurred in simulation run 62 for the weighted unscaled estimate of β_0. For the estimates of σ^2_{0k}, the differences were large in the simulation run 76 for the unweighted estimates, simulation run 62 for the weighted unscaled estimates and simulation run 54 for the weighted scaled 2 estimates.

We next determine what we would expect the results to be for each of the estimating models. The top row of Figure 3.4 contains the summary of the estimating model from Equation 3.11. When the estimated model matches the generating model, all of the estimation methods (PSHGR, RHS for all of unweighted, weighted unscaled, weighted scaled 1 and weighted scaled 2) have minimal bias for the β coefficients. It appears that the spread of the estimates using weighted unscaled weights is larger for the estimates of β_2

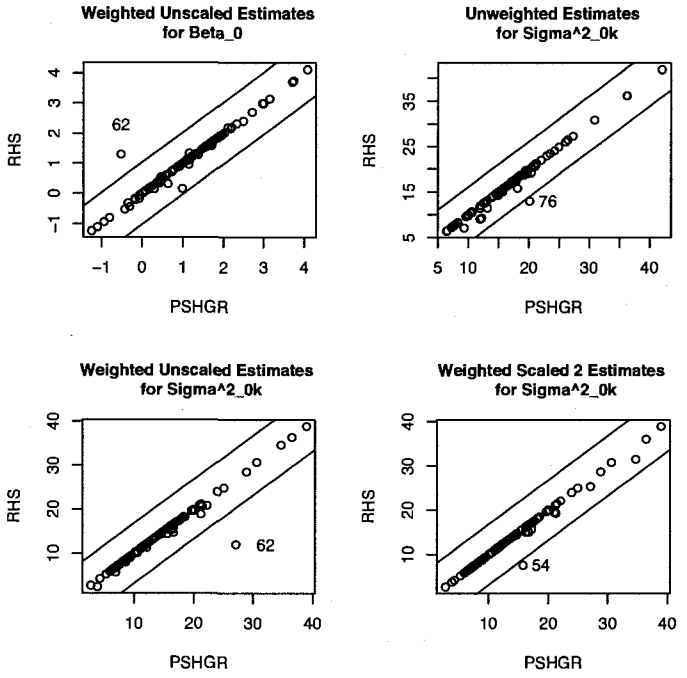

Figure 3.16: Comparison of PSHGR vs. RHS for Estimates from Equation 3.12

than the estimates using the other weighted methods. The unweighted estimates had a smaller spread than the weighted estimates. There is a small difference between the different weighting schemes in the estimation of the σ^2_{1k} parameter, but the differences are small compared to the differences in the σ^2_ϵ estimates. It appears that the σ^2_ϵ parameter follows the trends outlined in Section 2.4.2, with the weighted unscaled estimates having larger negative bias, the weighted scaled 1 estimates having smaller positive bias and the weighted scaled 2 estimates being in between them.

The second row of Figure 3.4 misspecifies the model by removing the random slope on x_{1k}, the cluster variable, and adds a random intercept. As expected, the estimation of

β_0, β_1 and β_2 are not affected by the misspecification. The random intercept includes the variation in the $U_{1k} \times x_{1k}$ variable. Recall that x_{1k} was generated as a normal random variable with mean 3 and variance 9 and U_{1k} was generated independently of x_{1k} as a normal random variable with mean 0 and variance 1. A quick simulation of 1000 sets of two simulated normal random variables set up similar to U_{1k} and x_{1k} provides variance of 18. The estimates in the figure are slightly lower (between 13.5 and 16) which follows the trend of the intercept variance having a negative bias when the ICC is large (see Section 2.4.2).

Results Description of Misspecification of Random Variables – Simulation Set 6

We want to flag if there are large differences between the PSHGR and RHS estimates for a given iteration. To do this, the standard deviation of the parameter estimate over the 100 iterations is obtained separately for the PSHGR and the RHS estimates. The smaller of these standard deviations is used as a threshold to flag "large" differences between PSHGR and RHS estimates. For each iteration, the difference between the PSHGR and the RHS estimates is compared to the threshold to identify estimates where the difference is greater than one standard deviation. Unless otherwise mentioned, the difference between the PSHGR and RHS estimates is less than the threshold. For the estimating model in Equation 3.14, simulation number 46 produced differences larger than the threshold in the PSHGR and RHS methods in 9 different estimates spanning all parameters and all weighting schemes. The plots to show these differences for each parameter and each scaling are not shown to conserve space. In addition, the weighted unscaled estimates of σ^2_{1k} also varied more than one threshold for simulation runs 46 and 56, see Figure 3.17. Of these differences, it was only the difference in the weighted unscaled estimate of σ^2_{1k} that was large enough to produce a difference in Figure 3.5. For the estimating model in Equation 3.15, there are four simulation runs whose PSHGR and RHS estimates differ by more than one threshold, as shown in Figure 3.18. These points correspond to simulation runs 55 and 94 for the weighted unscaled estimates of β_0 and simulation runs 39 and 94 for the weighted unscaled estimates of σ^2_{0k}.

We next determine what we would expect the results to be for each of the estimating

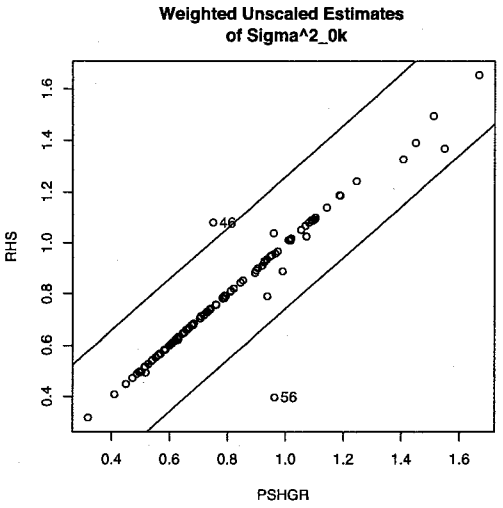

Figure 3.17: Comparison of PSHGR vs. RHS for Estimates from Equation 3.14

models. The top row of Figure 3.5 contains the summary of estimating model in Equation 3.14. As expected, when there is informative sampling of clusters based on the size of the random effect U_{1k}, the estimate of x_{1k} increases and the estimate of σ^2_{1k} decreases in the unweighted case. All of the weighted cases help to compensate for this informative sampling and the estimates are similar to those in Figure 3.4. It appears that the σ^2_{1k} parameter follows the trends of the random intercept outlined in Section 2.4.2. It appears that the σ^2_ϵ parameter follows the trends outlined in Section 2.4.2, with the weighted unscaled estimates having larger negative bias, the weighted scaled 1 estimates having smaller positive bias and the weighted scaled 2 estimates being in between them.

The second row of Figure 3.5 misspecifies the model by removing the random slope on x_{1k}, and adding a random intercept. As expected, the estimation of β_0, β_1, β_2 and σ^2_ϵ are not affected by the misspecification and have estimates similar to the top row,

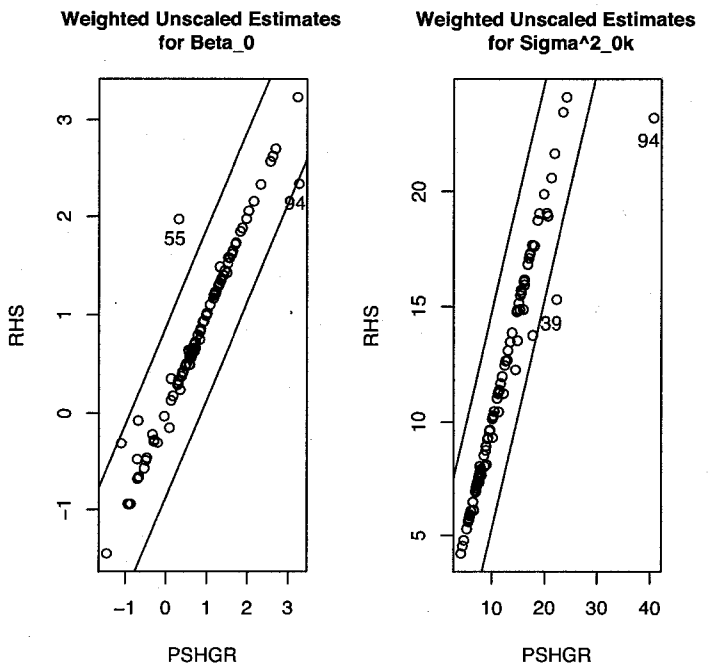

Figure 3.18: Comparison of PSHGR vs. RHS for Estimates from Equation 3.15

though the spread of β_0 and β_1 appear to be larger. The random intercept includes the variation in the $U_{1k} \times x_k$ variable. Recall that x_{1k} was generated as a normal random variable with mean 3 and variance 9 and U_{1k} was generated independently of x_{1k} as a normal random variable with mean 0 and variance 1. A quick simulation of 1000 sets of two simulated normal random variables set up similar to U_{1k} and x_{1k} provides variance around 19. The estimates in the figure are slightly lower which follows the trend of the intercept variance having a negative bias. As can be seen by comparing this Figure to Figure 3.4, the estimate of the unweighted σ^2_{01} is lower than the weighted estimates, which reflects the smaller variance in the sampled U_{1k} due to the informative sampling.

Results Description of Misspecification of Random Variables – Simulation Set 7

We want to flag if there are large differences between the PSHGR and RHS estimates for a given iteration. To do this, the standard deviation of the parameter estimate over the 100 iterations is obtained separately for the PSHGR and the RHS estimates. The smaller of these standard deviations is used as a threshold to flag "large" differences between PSHGR and RHS estimates. For each iteration, the difference between the PSHGR and the RHS estimates is compared to the threshold to identify estimates where the difference is greater than one standard deviation. Unless otherwise mentioned, the difference between the PSHGR and RHS estimates is less than the threshold. For the estimating model in Equation 3.17, simulation set 23 produced differences larger than the threshold in the PSHGR and RHS methods in 12 different estimates spanning all parameters and all weighting schemes. The plots to show these differences for each parameter and each scaling are not shown to conserve space. For the estimating model in Equation 3.18, the scaled 1 estimates of σ_ϵ^2 produced differences between PSHGR and RHS greater than the threshold in simulation runs 46 and 56, as seen in Figure 3.19.

We next determine what we would expect the results to be for each of the estimating models. The top row of Figure 3.6 contains the summary of estimating model from Equation 3.17. When the estimated model matches the generating model, all of the estimation methods (PSHGR, RHS for all of unweighted, weighted unscaled, weighted scaled 1 and weighted scaled 2) have minimal bias and comparable quantiles. The exception to this is that the weighted unscaled estimates appear to have larger spread for the β_0 and β_1 param-

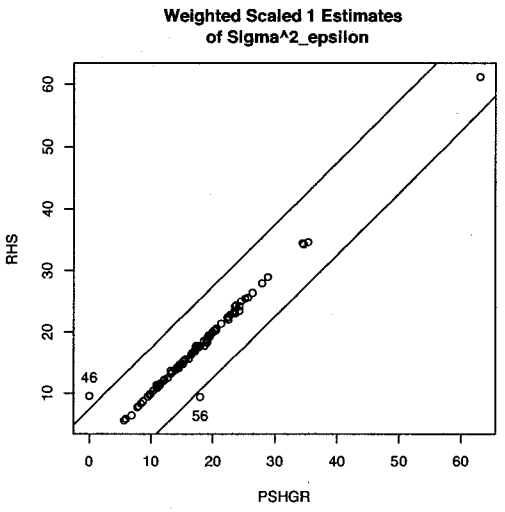

Figure 3.19: Comparison of PSHGR vs. RHS for Estimates from Equation 3.18

eters. The estimates of the σ^2_{2k} parameter appear to be quite similar, with the exception of the unweighted estimates, that have slightly less bias. The σ^2_ϵ parameter follows the trends outlined in Section 2.4.2, with the weighted unscaled estimates having larger negative bias, the weighted scaled 1 estimates having smaller positive bias and the weighted scaled 2 estimates being in between them.

The second row of Figure 3.6 misspecifies the model by removing the random slope on x_{2ik}, the unit variable, and adds a random intercept. As expected, the estimation of β_0, β_1 and β_2 are not affected by the misspecification. The random intercept includes the variation in the $U_{2k}x_{2ik}$ variable. Recall that x_{2ik} was generated as a normal random variable with mean 1 and variance 25 and U_{2k} was generated independently of x_{2ik} as a normal random variable with mean 0 and variance 0.8. We would expect a portion of the variance to go into the estimate of σ^2_ϵ and a portion to go into the σ^2_{0k}. If you condition first on the values of U_{2k},

the random error variance for that cluster will increase by $U_{2k}^2 * \text{Var}(x_{2ik})$, approximately $0.8^2 * 25 = 16$. Alternatively, if we condition on x_{2ik} then the random intercept variance will increase by roughly $\bar{x}_{2ik}^2 \text{Var}(U_{2k})$, approximately $1^2 * 0.8 = 0.8$. That would provide a random intercept variance of approximately 0.8+0.8=1.6, and a random error variance of approximately 0.5+16=16.5. The simulation results are consistent with these results.

Results Description of Misspecification of Random Variables – Simulation Set 8

We want to flag if there are large differences between the PSHGR and RHS estimates for a given iteration. To do this, the standard deviation of the parameter estimate over the 100 iterations is obtained separately for the PSHGR and the RHS estimates. The smaller of these standard deviations is used as a threshold to flag "large" differences between PSHGR and RHS estimates. For each iteration, the difference between the PSHGR and the RHS estimates is compared to the threshold to identify estimates where the difference is greater than one standard deviation. In this simulation set, there were no differences greater than the threshold.

We next determine what we would expect the results to be for each of the estimating models. The top row of Figure 3.7 contains the summary of estimating model from Equation 3.20. As expected, when there is informative sampling of units based on the size of the random effect U_{2k}, the estimate of x_{2ik} increases and the estimate of σ^2_{2k} decreases in the unweighted case. All of the weighted cases help to compensate for this informative sampling and the estimates are similar to those in Figure 3.6. The weighted σ^2_{2k} estimates all have similar point estimates and ranges. The σ^2_ϵ parameter follows the trends outlined in Section 2.4.2, with the weighted unscaled estimates having larger negative bias, the weighted scaled 1 estimates having smaller non-negative bias and the weighted scaled 2 estimates being in between them.

The second row of Figure 3.6 misspecifies the model by removing the random slope on x_{2ik}, the unit variable, and adds a random intercept. As expected, the estimation of

β_0, β_1 and β_2 are not affected by the misspecification. The random intercept includes the variation in the $U_{2k}x_{2ik}$ variable. Recall that x_{2ik} was generated as a normal random variable with mean 1 and variance 25 and U_{2k} was generated independently of x_{2ik} as a normal random variable with mean 0 and variance 0.8. We would expect a portion of the variance to go into the estimate of σ^2_ϵ and a portion to go into the $\sigma^2_0 k$. If you condition first on the values of U_{2k}, the random error variance for that cluster will increase by $U^2_{2k} * \text{Var}(x_{2ik})$, approximately $0.8^2 * 25 = 16$. Alternatively, if we condition on x_{2ik} then the variance will be roughly $\bar{x}^2_{2ik}\text{Var}(U_{2k})$, approximately $1^2 * 0.8 = 0.8$. That would provide a random intercept variance of approximately .8+16=16.8, and a random error variance of approximately 0.5+0.8=1.3. The simulation supports these conclusions.

Results Description of Misspecification of Stratification Layers – Simulation Set 9

We want to flag if there are large differences between the PSHGR and RHS estimates for a given iteration. To do this, the standard deviation of the parameter estimate over the 100 iterations is obtained separately for the PSHGR and the RHS estimates. The smaller of these standard deviations is used as a threshold to flag "large" differences between PSHGR and RHS estimates. For each iteration, the difference between the PSHGR and the RHS estimates is compared to the threshold to identify estimates where the difference is greater than one standard deviation. Unless otherwise mentioned, the difference between the PSHGR and RHS estimates is less than the threshold. For the estimating model in Equation 3.23, there were a number of simulation runs that produced estimates the unweighted estimates of σ^2_{02k} where the differences between PSHGR and RHS greater than the threshold, as shown in Figure 3.20. These include simulation run 4 for the unweighted estimates, simulation runs 16 and 81 for the weighted unscaled estimates and simulation runs 19, 92 and 94 for the weighted scaled 1 estimates. The differences in the unweighted and weighted unscaled estimates are too small to be seen in Figure 3.8. However, the difference in the weighted scaled 1 estimates is seen due to the extreme values of the RHS estimates. In addition, the PSHGR and RHS estimates of the covariance term $\sigma^2_{01k.02k}$ were quite different, as seen in Figure 3.21. Further investigation is needed to better understand why the spread of the estimates are so different. The PSHGR covariance estimates are all very close to zero (less than 10^{-16} in absolute value), whereas the RHS estimates vary between approximately 3 and -3. However, the RHS weighted unscaled estimates have a few

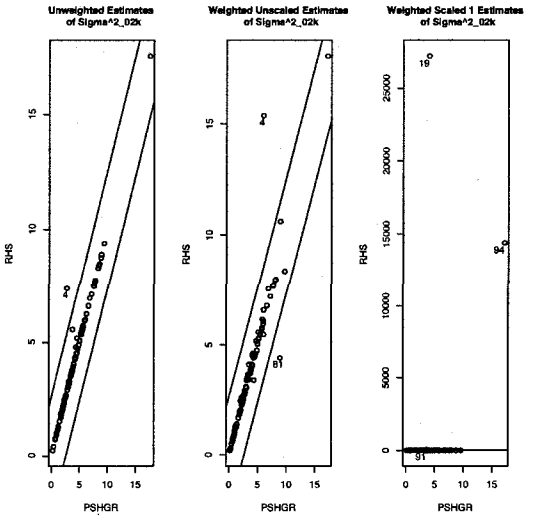

Figure 3.20: Comparison of PSHGR vs. RHS for Estimates from Equation 3.23

large outliers. These differences between the RHS and PSHGR estimates are clear in Figure 3.8. Note that the weighted scaled 1 estimates of $\sigma^2_{01k.02k}$ also follow a different pattern than the other estimates because of the extreme values of the RHS estimates. For the estimating model in Equation 3.24, the PSHGR and RHS weighted unscaled estimates of σ^2_{0k} for simulation run 16 are larger than the threshold, as are the estimates from simulation run 64 for the weighted unscaled estimates, as seen in Figure 3.22. These differences are not large enough to be seen in Figure 3.8. For the estimating model in Equation 3.26, the PSHGR and RHS weighted unscaled estimates of σ^2_{0k} for simulation runs 16, 73 and 77 are larger than the threshold, as are the estimates from simulation run 27 for the weighted scaled 2 estimates, as seen in Figure 3.23. These differences are not large enough to be seen in Figure 3.8.

We next determine what we would expect the results to be for each of the estimating

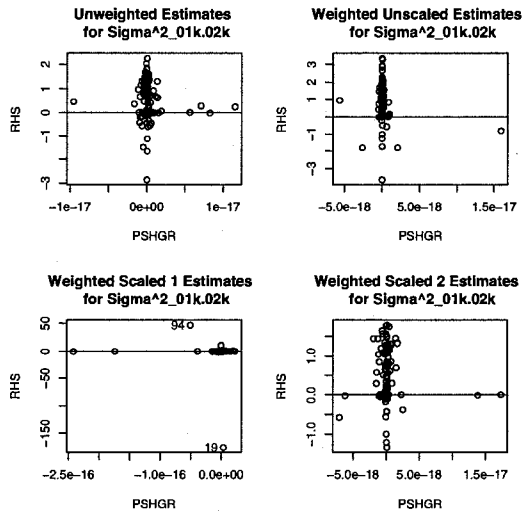

Figure 3.21: Comparison of PSHGR vs. RHS for Estimates of $\sigma^2_{01k.02k}$ from Equation 3.23

models. In Figure 3.8, the first row shows the summary from the estimating model in Equation 3.23. There are two fixed effects in this regression and all estimation methods perform well. Besides the differences in the estimates between PSHGR and RHS described above, there is nothing else notable regarding the variance components. Finally, when the generating model equals the estimating model, the estimates of σ^2_ϵ follow the same trends as the previous simulations.

The second row of Figure 3.8 shows a summary of the results from the estimating model in Equation 3.24. This model is misspecified because the stratified/clustered design is estimated as a clustered design. Recall is no informative sampling. In this model, the two strata are being estimated as one. Since the number of elements in each strata are roughly equal, I would expect that the estimated intercept would be the average of the intercept of the two strata, in this case $(-3+5)/2 = 1$, and the graph supports this. The estimate

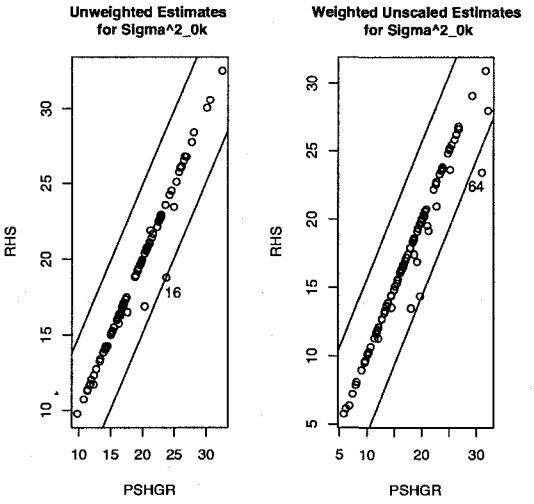

Figure 3.22: Comparison of PSHGR vs. RHS for Estimates from Equation 3.24

of σ_ϵ^2 is about the true value of 0.5 as the variance within each cluster should remain unchanged. The random intercept should pick up the variance associated with dropping the two strata. Note that roughly 50 sampled elements in stratum 1 have an intercept of 5, and the roughly 50 sampled elements in stratum 2 have an intercept of -3. The variance of this will be roughly $\frac{1}{100}(\sum_{i=1}^{50}(5-1)^2 + \sum_{i=1}^{50}(-3-1)^2) = 16$. The variance of 16 assumes that each strata has a fixed effect intercept. Because there are random intercepts within each stratum, the variance due to the random intercepts needs to be taken into account by increasing 16 by $\text{Var}(U_{s1} + U_{s2})/2) = 6/4 = 1.5$ to 17.5. This is consistent with the figure.

The third row of Figure 3.8 contains a summary from the estimated model in Equation 3.26. The generating model is in Equation 3.25. This is the same as the other two simulations in this simulation set, except that the sampling design informatively sampled clusters based on the size of their random effects. When comparing the estimates of β_0, both the

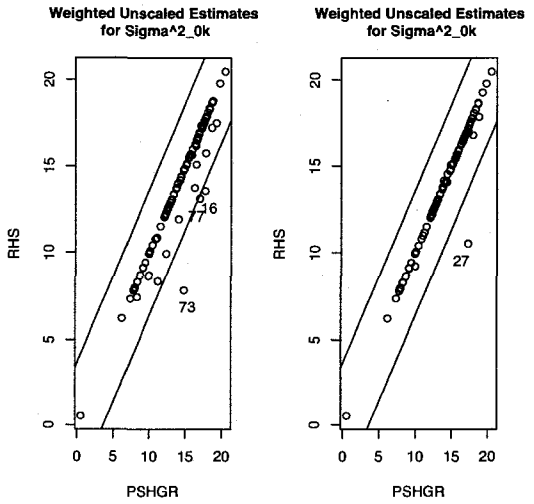

Figure 3.23: Comparison of PSHGR vs. RHS for Estimates from Equation 3.26

weighted and the unweighted estimates from Equation 3.26 are larger than those in Equation 3.24. In addition, the unweighted estimates from the estimated model in Equation 3.26 are larger compared to the weighted estimates than those from the estimated model in Equation 3.24. In addition, all of the estimates of σ^2_{0k} are smaller than the estimates from the estimating model in Equation 3.24. In addition the unweighted estimates of σ^2_{0k} are smaller than the weighted estimates, especially when compared to the estimates of σ^2_{0k} from the estimated model in Equation 3.24.

Results Description of Misspecification of Stratification Layers - Simulation Set 10

We want to flag if there are large differences between the PSHGR and RHS estimates for a given iteration. To do this, the standard deviation of the parameter estimate over the 100 iterations is obtained separately for the PSHGR and the RHS estimates. The smaller of these standard deviations is used as a threshold to flag "large" differences between PSHGR and RHS estimates. For each iteration, the difference between the PSHGR and the RHS estimates is compared to the threshold to identify estimates where the difference is greater than one standard deviation. Unless otherwise mentioned, the difference between the PSHGR and RHS estimates is less than the threshold. For the estimating model in Equation 3.28, there are simulation runs that produced differences between PSHGR and RHS greater than the threshold. For the estimating model in Equation 3.29, there are differences between PSHGR and RHS in the unweighted and weighted scaled 1 estimates of σ_ϵ^2, as shown in Figure 3.24. These differences are from simulation runs 5, 33, 65, 80 and 97 for the unweighted estimates and simulation runs 16, 20, 30, 35, 44, 51, 53, 57, 58, 60, 63, 64 and 95. For the weighted scaled 1 estimates of σ_ϵ^2, it is clear that most of the differences are caused when PSHGR is estimating the parameter near 0, whereas RHS is estimating the parameter between 14 and 21. I suspect this is a problem with the PSHGR computations. For the estimating model in Equation 3.31, there are also differences between PSHGR and RHS estimates. Figures 3.25 and 3.26 show the differences between PSHGR and RHS in the β_0, σ_{0k}^2 and σ_ϵ^2 parameters. Figure 3.25 shows that the there is a large difference between the weighted unscaled estimates of β_0 for simulation run 40, between the weighted

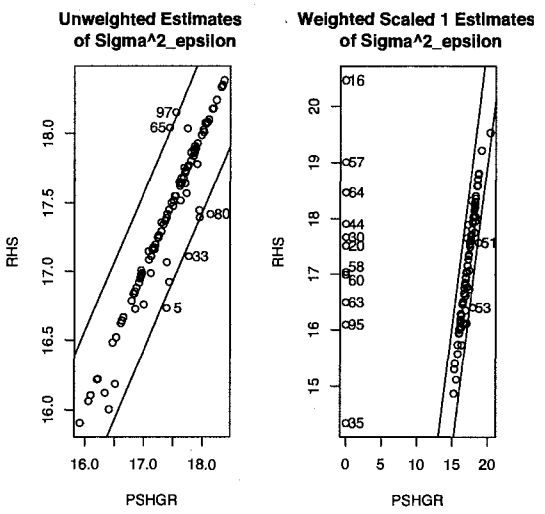

Figure 3.24: Comparison of PSHGR vs. RHS for Estimates from Equation 3.29

scaled 2 estimates of σ^2_{0k} for simulation run 57, between the weighted unscaled estimates of σ^2_{0k} for runs 40 and 26, and between the weighted unscaled estimates of σ^2_ϵ for run 40. Figure 3.26 shows the difference between PSHGR and RHS in the weighted scaled 1 estimates of σ^2_ϵ. Similar to Figure 3.24, the PSHGR method has many estimates near 0, whereas the same data produced estimates between 15 and 20 for RHS. The problematic simulation runs were 3, 11, 22, 37, 43, 52, 64, 69, 76, 80, 90, 92, and 97. Again, it is clear that most of the differences are caused when PSHGR is estimating the parameter near 0, whereas RHS is estimating the parameter between 14 and 21. I suspect this is a problem with the PSHGR computations.

We next determine what we would expect the results to be for each of the estimating models. In Figure 3.9, the first row shows the estimates of the parameters when the estimating model from Equation 3.28 matches the generating model. All of the weighting

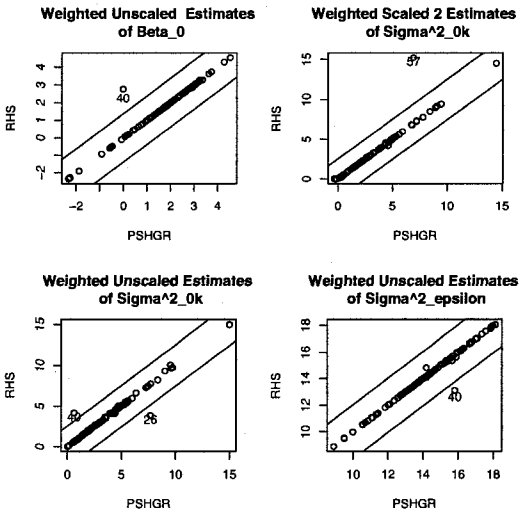

Figure 3.25: Comparison of PSHGR vs. RHS for Estimates from Equation 3.31

methods estimate the β parameter well. The unweighted estimates have a smaller spread. The spread for the weighted unscaled estimation for β_1 appears to be wider than the other weighted methods. The estimate of σ^2_{0k} appears to be the similar across different weighting methods, likely due to the higher intra-class correlation. The pattern of the estimates for σ^2_ϵ is the same as in previous simulations.

The second row of Figure 3.9 misspecifies the model by removing the stratification, so that the stratified/clustered design is estimated as a clustered design, as detailed in Equation 3.29. In this second row, the clusters are sampled proportional to an independent random variable (non-informatively). Here, I would expect the variance of the random intercept to remain the same, and the random error term variance, σ^2_ϵ to absorb the variance from not including the stratification in the model. Note that roughly half sampled elements in a cluster are in stratum 1 with an intercept of 5, and the roughly half sampled elements in

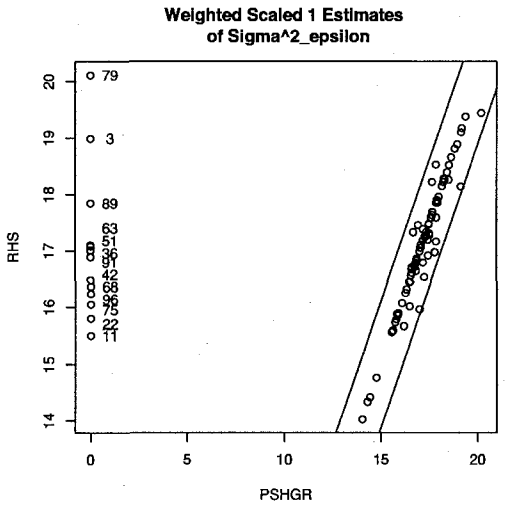

Figure 3.26: Comparison of PSHGR vs. RHS for Estimates from Equation 3.31 (cont)

a cluster are in stratum 2 with an intercept of -3. If n_k is the number of elements in cluster k, then the variance of the error term will be roughly $\frac{1}{n_k}(\sum_{i=1}^{n_k/2}(5-1)^2+\sum_{i=1}^{n_k/2}(-3-1)^2) = 16$. Adding this to the original random error of 0.5 gives an estimated value of σ_ϵ^2 of about 16.5, as seen in the figure. Because the intra-class correlation is smaller now due to the increase in the random error variance, the estimates of σ_{0k}^2 are exhibiting the behavior of the previous simulations with a low intra-class correlation.

The third row in Figure 3.9 sampled the clusters informatively, porportional to the size of the random effect (U_{0k}), as detailed in Equation 3.31. Because of this, the estimate of the random intercept is larger in the unweighted case and the estimate of the variance of the random intercept is smaller The smaller variance in the unweighted estimate can be seen by comparing the unweighted estimate of the random intercept in the second row of Figure 3.9 with the unweighted estimate of the random intercept of the third row of the

same figure. These are corrected with the weighted estimates. The estimate of σ_ϵ^2 remains unchanged, as expected.

Results Description of Misspecification of Stratification Layers - Simulation Set 11

We want to flag if there are large differences between the PSHGR and RHS estimates for a given iteration. To do this, the standard deviation of the parameter estimate over the 100 iterations is obtained separately for the PSHGR and the RHS estimates. The smaller of these standard deviations is used as a threshold to flag "large" differences between PSHGR and RHS estimates. For each iteration, the difference between the PSHGR and the RHS estimates is compared to the threshold to identify estimates where the difference is greater than one standard deviation. Unless otherwise mentioned, the difference between the PSHGR and RHS estimates is less than the threshold. Figures 3.27, 3.28 and 3.29 contain graphs of the estimates from the simulation runs whose difference between the PSHGR and RHS estimates is larger than the thresholds from the estimating model in Equation 3.33. Note that simulation run 10 did not converge for the RHS weighted unscaled estimates. From Figure 3.27, we see that PSHGR and RHS methods differed for simulation run 71 in the weighted scaled 1 estimates of β_0, β_1 and σ^2_{01k}. The effect of the large difference from run 71 can be seen in Figure 3.10 in the difference between the means of the RHS and PSHGR scaled 1 estimates of β_0 and σ^2_{01k}. In addition, the PSHGR and RHS weighted unscaled estimates of β_1 differed by more than the threshold in simulation run 60.

Figure 3.28 contains the graphs of PSHGR vs. RHS estimates for the σ^2_{02k} parameter. All of the weighting methods contained simulation runs that produced large differences between the PSHGR and RHS estimates. For the unweighted estimates, simulation runs 17 and 23 produced large differences. Note that in Figure 3.10, the mean of RHS unweighted

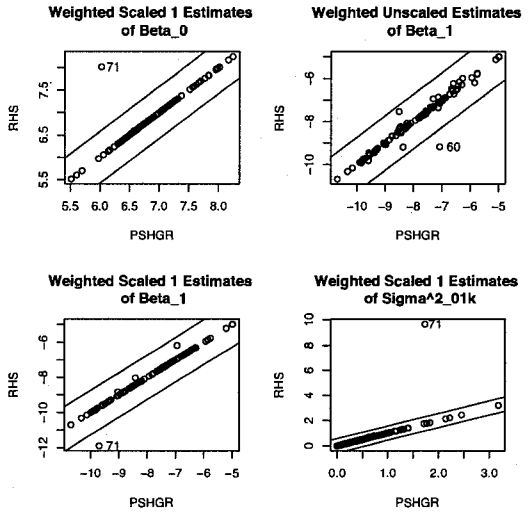

Figure 3.27: Comparison of PSHGR vs. RHS for Estimates from Equation 3.33

estimates of σ^2_{02k} is larger than the spread of the 0.025 to 0.975 quantiles. This is due to the large value from simulation run 17. For the weighted unscaled estimates, simulation runs 11, 13, 26, 32, 36, 38, 51, 60, 65, 89, and 97 produced large differences. In Figure 3.10, these large differences are reflected as a much larger spread and mean for the RHS weighted unscaled estimates than for the PSHGR weighted unscaled estimates. For the weighted scaled 1 estimates, simulation runs 2 and 25 produced large estimates. Finally, for the weighted scaled 2 estimates, simulation runs 13, 45 and 63 produced large differences. These differences are shown in Figure 3.10 in that the mean of the RHS estimate is not on the graph. The large value(over 80,000) from simulation run 13 for RHS causes the RHS mean to be larger than the scale printed in the figure.

Figure 3.29 contains the graphs of PSHGR vs. RHS estimates for the $\sigma^2_{01k.02k}$ parameter. This trend is similar to the estimated covariance term from Equation 3.23 seen in Figure

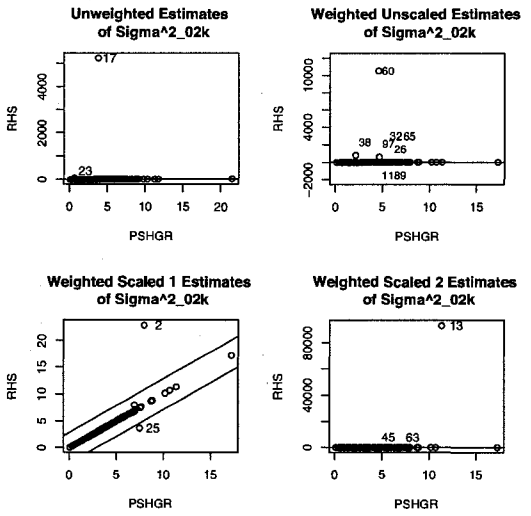

Figure 3.28: Comparison of PSHGR vs. RHS for Estimates of σ^2_{02k} from Equation 3.33

3.21. The PSHGR estimates are showing a small amount of variablility (note that the scales on the x-axis are no larger than $\pm 4 \times 10^{-17}$). The RHS scales are roughly ± 3, except for the weighted scaled 1 estimates where the RHS has some large estimates, around 50 and -170 and the weighted scaled 2 estimates about 250. In Figure 3.10 it is clear that the spread of the PSHGR estimates is smaller than the RHS estimates. The larger estimates of the RHS scaled 1 estimates is reflected in a larger spread in the figure. In addition, the large value (about 250) of the RHS weighted scaled 2 estimate is causing the mean to be large in Figure 3.10.

Figure 3.30 contains the graphs of PSHGR vs. RHS estimates for the σ^2_{0k} parameter from the estimating model in Equation 3.34. In general, these simulation show many differences between PSHGR and RHS in this parameter, except there are no differences for the weighted scaled 1 estimates. For the unweighted estimates, simulation runs 2, 3, 16,

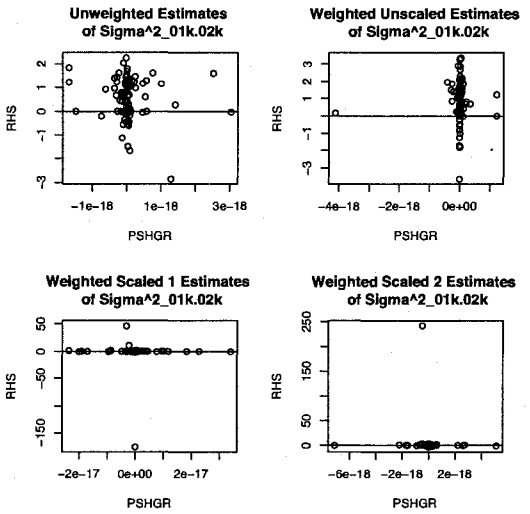

Figure 3.29: Comparison of PSHGR vs. RHS for Estimates of $\sigma^2_{01k.02k}$ from Equation 3.33

21, 28, 35, 39, 53, 59, 67, 74, 75, and 87 produced estimates with differences larger than the threshold. For the weighted unscaled estimates, simulation runs 8, 27, 48, 52, 53, 62, 76, 79, 82, 84, and 93 produced estimates with differences larger than the threshold. For the weighted scaled 2 estimates, simulation runs 32, 68, 71, 78, and 89 produced estimates with differences larger than the threshold. These differences can be seen in Figure 3.10 in the comparison of the PSHGR and RHS unweighted, weighted unscaled and weighted scaled 2 estimates.

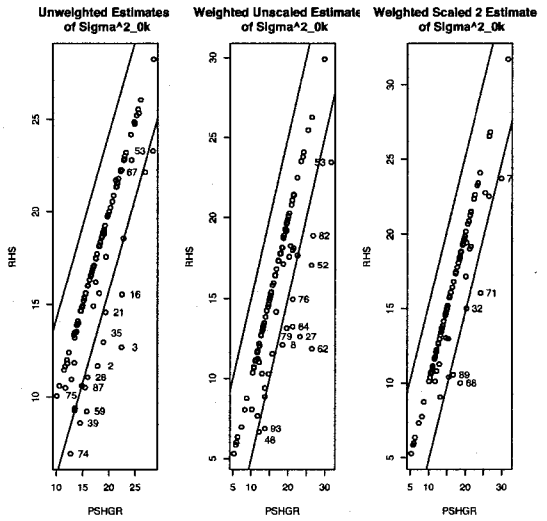

Figure 3.30: Comparison of PSHGR vs. RHS for Estimates from Equation 3.34

There are many issues with the estimation from the estimated model in Equation 3.35. The PSHGR method produces estimates in only 75 of the 100 simulation runs. This is mostly due to not being able to invert the Ω matrix or the A matrix in the computation of $V-1$ from Equations 2.34, 2.35, 2.36 and 2.37. This needs to be further investigated. The simulation runs that contained computation problems are 4, 6, 7, 19, 20, 23, 26, 30, 32, 34, 36, 41, 43, 45, 58, 63, 65, 68, 72, 74, 80, 81, 86, 90, and 94. In addition, there are a number of PSHGR runs that did not converge within 500 iterations for the weighted scaled 1 estimates, including runs 12, 14, 15, 21, 27, 31, 54, 55, 62, 67, 77, 83, 91, 93, 97, 98, 99, and 100. The RHS method did not converge for simulation run 6 for the scaled 1 estimates and for simulation run 71 for the scaled 2 estimates. As can be seen in Figures 3.31 to 3.36, the estimation from this model produces many differences between PSHGR and RHS.

Figure 3.31 contains the graphs of PSHGR vs. RHS estimates for the estimate of β_0. The weighted unscaled estimates produced differences between PSHGR and RHS larger than the threshold for simulation run 75. The weighted scaled 1 estimates produce differences between PSHGR and RHS larger than the threshold for simulation runs 28, 50, and 75. The weighted scaled 2 estimates produced differences between PSHGR and RHS larger than the threshold for simulation run 37. The differences between the weighted scaled 1 estimates of PSHGR and RHS can be seen in Figure 3.10 as the PSHGR 0.025 quantile and mean are lower than the corresponding RHS values. The other differences are too small to notice on the figure.

Figure 3.32 contains the graphs of PSHGR vs. RHS estimates for the estimate of

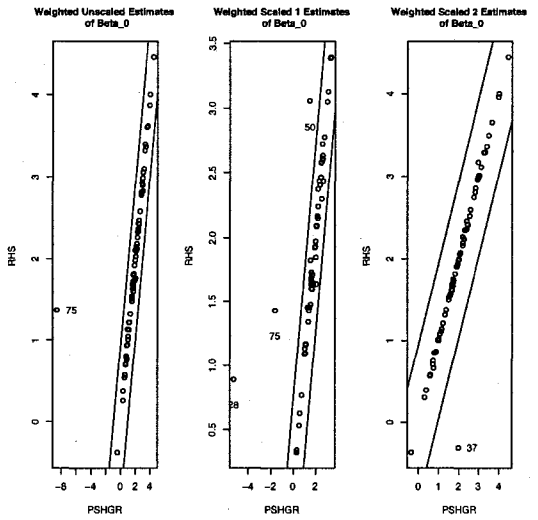

Figure 3.31: Comparison of PSHGR vs. RHS for Estimates of β_0 from Equation 3.35

β_1. The weighted unscaled estimates produced differences between PSHGR and RHS larger than the threshold for simulation run 75. The weighted scaled 1 estimates produce differences between PSHGR and RHS larger than the threshold for simulation runs 28 and 75. The weighted scaled 2 estimates produced differences between PSHGR and RHS larger than the threshold for simulation runs 37 and 62. The differences are reflected in Figure 3.10 by PSHGR having a larger quantile than RHS for the scaled 1 estimate of β_1 and PSHGR having a smaller 0.025 quantile than RHS for the weighted unscaled estimates of β_1. The other differences are too small to notice on the figure.

Figure 3.33 contains the graphs of PSHGR vs. RHS estimates for the estimate of σ^2_{01k}. The unweighted estimates produced differences between PSHGR and RHS larger than the threshold for simulation run 2. The weighted unscaled estimates produce differences between PSHGR and RHS larger than the threshold for simulation runs 2, 3, 10, 15, 16,

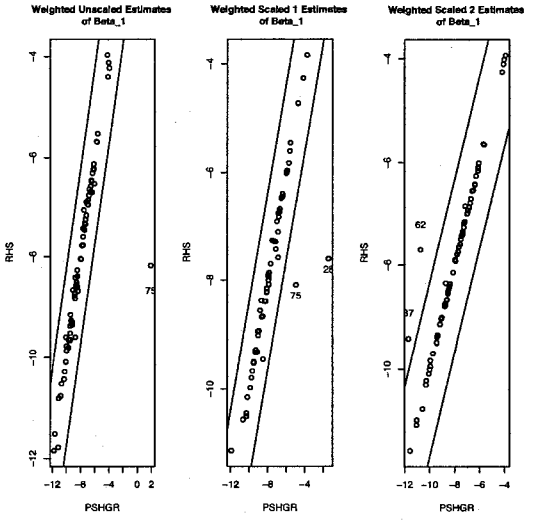

Figure 3.32: Comparison of PSHGR vs. RHS for Estimates of β_1 from Equation 3.35

21, 22, 25, 39, 40, 50, 53, 79, 84, 85, and 97. The weighted scaled 2 estimates produced differences between PSHGR and RHS larger than the threshold for simulation runs 3, 10, 13 and 47. The other differences are too small to notice on the figure.

Figure 3.34 contains the graphs of PSHGR vs. RHS estimates for the estimate of σ^2_{02k}. The unweighted estimates produced differences between PSHGR and RHS larger than the threshold for simulation run 71. The weighted unscaled estimates produce differences between PSHGR and RHS larger than the threshold for simulation runs 9, 28 and 57. The weighted scaled 2 estimates produced differences between PSHGR and RHS larger than the threshold for simulation runs 37 and 42. These differences are reflected in Figure 3.10 because the RHS weighted scaled 2 mean is so large (due to the simulation run 37 having an estimate of 2500) that it is not printed on the plot for σ^2_{01k}. These differences are reflected in Figure 3.10 by RHS having a larger 0.975 quantile for the weighted unscaled estimates

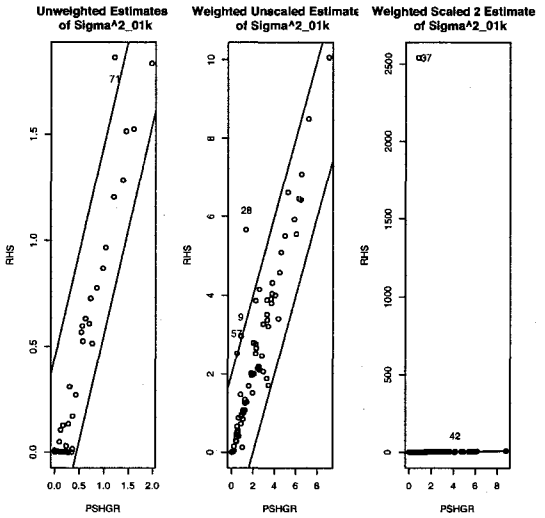

Figure 3.33: Comparison of PSHGR vs. RHS for Estimates of σ^2_{01k} from Equation 3.35

than PSHGR. Also, the RHS simulation runs 3, 10, 13 and 47 cause the RHS 0.975 quantile to be larger than the PSHGR corresponding quantile for the weighted scaled 2 estimates. The mean of the RHS weighted scaled 2 estimate of σ^2_{02k} is printed off of the scale of the graph on Figure 3.10. The other differences are too small to notice on the figure.

Figure 3.35 contains the graphs of PSHGR vs. RHS estimates for the estimate of $\sigma^2_{01k.02k}$. In this figure, we see the same trends as we did in Figures 3.21 and 3.29. The variation in the PSHGR estimates is very small, with the largest variation being approximately 2^{-13}. The RHS estimates have more spread, with the weighted scaled 2 estimates containing two large estimates around 3000 and 6000. This pattern should be looked into further. These differences are reflected in Figure 3.10 by the small ranges for the PSHGR estimates and the larger ranges for the RHS weighted unscaled and weighted scaled 2 estimates. In addition, the mean of the RHS weighted scaled 2 estimates is so large (due to

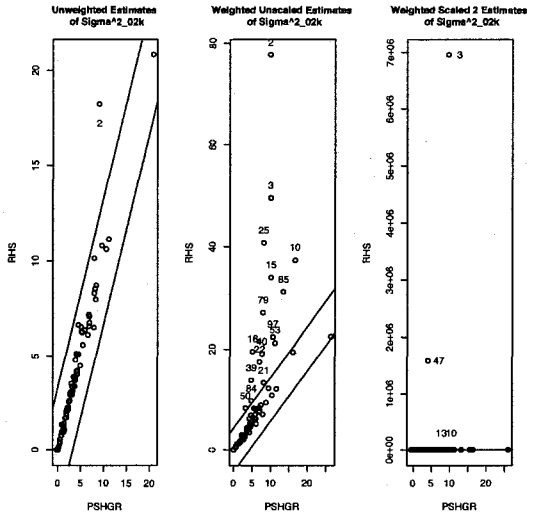

Figure 3.34: Comparison of PSHGR vs. RHS for Estimates of σ^2_{02k} from Equation 3.35

the estimates of 6000 and 3000) that it is not printed on the range of the graph. The other differences are too small to notice on the figure.

Figure 3.36 contains the graphs of PSHGR vs. RHS estimates for the estimate of σ^2_ϵ. The unweighted estimates produced differences between PSHGR and RHS larger than the threshold for simulation runs 2, 16, 35, 37, 59 and 71. The weighted unscaled estimates produce differences between PSHGR and RHS larger than the threshold for simulation run 75. The weighted scaled 1 estimates produced differences between PSHGR and RHS larger than the threshold for simulation runs 28, 37, 40, 50, 73, 75 and 76. The differences are reflected in Figure 3.10 by small value of the 0.025 PSHGR quantile of the weighted scaled 1 estimates. The other differences are too small to notice on the figure.

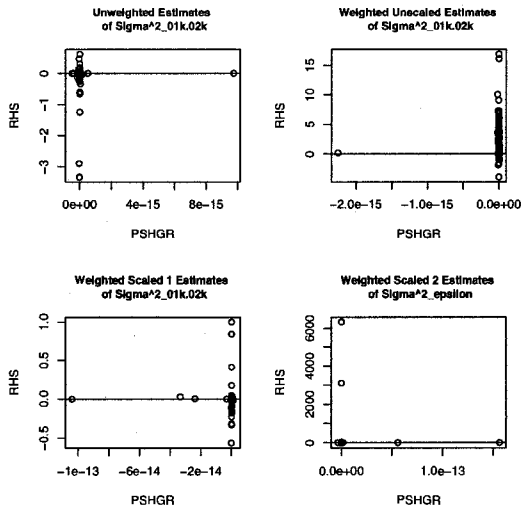

Figure 3.35: Comparison of PSHGR vs. RHS for Estimates of $\sigma^2_{01k.02k}$ from Equation 3.35

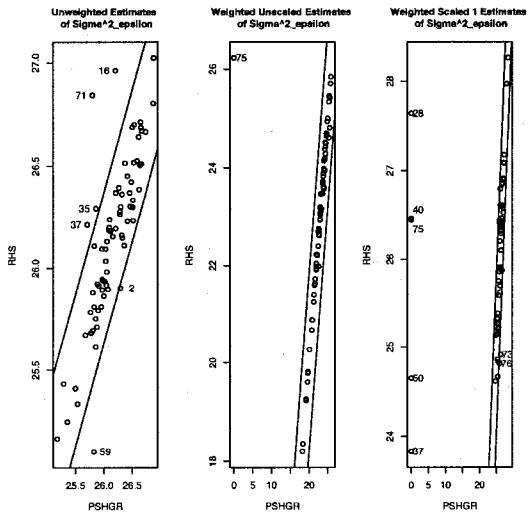

Figure 3.36: Comparison of PSHGR vs. RHS for Estimates of σ^2_ϵ from Equation 3.35

Figure 3.37 contains the estimates of σ^2_{0k} from the estimated model in Equation 3.36. The weighted unscaled estimates produced differences between PSHGR and RHS larger than the threshold for simulation runs 30 and 65. These differences are too small to notice on the Figure 3.10.

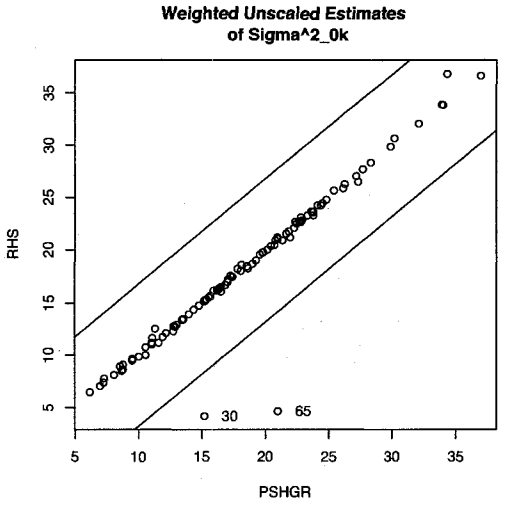

Figure 3.37: Comparison of PSHGR vs. RHS for Estimates of σ^2_{0k} from Equation 3.36

We next determine what we would expect the results to be for each of the estimating models. In Figure 3.10, the first row shows the estimates of the parameters when the estimating model matches the generating model. Generally, the estimation does well with the exception of the large estimates of RHS outlined above.

The second row of Figure 3.10 misspecifies the model by removing the top level of stratification, so that the stratified/clustered/stratified design is estimated as a clustered/stratified design. In this second row, the clusters are sampled proportional to an independent random variable (non-informatively). In this model, the two top level strata are being estimated

as one. Since the number of elements in each strata are roughly equal, I would expect that the estimated intercept would be the average of the intercept of the two strata, in this case $(-3+5)/2 = 1$. However, now the reference point for the intercept is the lower level stratum 1, that has an intercept of two. Thus, the intercept is now 1+2=3, and the graph supports this. The second level stratum level two coefficient has not changed. The estimate of σ^2_ϵ is about the true value of 0.5 as the variance within each cluster should remain unchanged. The random intercept should pick up the variance associated with dropping the two strata. Note that roughly 50 sampled elements in stratum 1 have an intercept of 5, and the roughly 50 sampled elements in stratum 2 have an intercept of -3. The variance of this will be roughly $\frac{1}{100}(\sum_{i=1}^{50}(5-1)^2 + \sum_{i=1}^{50}(-3-1)^2) = 16$. In addition, there is the variance from the random intercepts. Here, the random intercepts are $\text{var}((U_{s1}+U_{s2})/2) = 6/4 = 1.5$. This would lead to the overall random intercept with a variance of 16+1.5=17.5. This is consistent with the figure.

The third row of Figure 3.10 misspecifies the model by removing the second level of stratification, so that the stratified/clustered/stratified design is estimated as a stratified/clustered design. In this second row, the clusters are sampled proportional to an independent random variable (non-informatively). The intercept now represents the top level of stratification (averaged over the bottom level of stratification). The average of the bottom level of stratification is $(2-8)/2 = -3$. Thus, the intercept should be $5-3 = 2$, as shown in the graph. Note that the variance components for the RHS weighted scaled 2 estimation method have large spreads. This is due to two simulations creating large outliers for these estimates. I would expect the variance of the random intercept to remain the same,

and the random error term variance , σ_ϵ^2 to absorb the variance from not including the stratification in the model. Note that roughly half sampled elements in a cluster are in the first lower level stratum with an intercept of 2, and the roughly half sampled elements in a cluster are in the second lower level stratum with an intercept of -8. If n_{ks} is the number of elements in cluster k where S2=1 (or S2=2, as the strata are roughly equally sized), then the variance of the error term will be roughly $\frac{1}{2*n_{ks}}(\sum_{i=1}^{n_{ks}}(2+3)^2 + \sum_{i=1}^{n_k}(-8+3)^2) = 25$. Adding this to the original random error of 0.5 gives an estimated value of σ_ϵ^2 of about 25.5, as seen in the figure.

The fourth row of Figure 3.10 misspecifies the model by removing the all levels of stratification, so that the stratified/clustered/stratified design is estimated as a clustered design. In this second row, the clusters are sampled proportional to an independent random variable (non-informatively). The intercept now represents the average across all strata. We know that s1=1 has an intercept of 5, s1=2 has an intercept of -3, S2=1 has an intercept of 2 and S2=2 has an intercept of -8. Averaging these (as they all have roughly the same number of people) provides a grand intercept of -1, as indicated by the figure. Removing the lower level of stratification (the S2 level) will increase the estimate of σ_ϵ^2. The increase will be by 25, as indicated in the description in the above paragraph. Thus, the estimated σ_ϵ^2 should be 25+0.5 = 25.5, which is supported by the figure. In addition, the variance induced by removing the top level of stratification is put into the random intercept. As described in the previous two paragraphs, the variance of the random intercept should be about 16.5, as represented in the figure.

Results Description of Misspecification of Clustering Layers - Simulation Set 12

We want to flag if there are large differences between the PSHGR and RHS estimates for a given iteration. To do this, the standard deviation of the parameter estimate over the 100 iterations is obtained separately for the PSHGR and the RHS estimates. The smaller of these standard deviations is used as a threshold to flag "large" differences between PSHGR and RHS estimates. For each iteration, the difference between the PSHGR and the RHS estimates is compared to the threshold to identify estimates where the difference is greater than one standard deviation. Unless otherwise mentioned, the difference between the PSHGR and RHS estimates is less than the threshold. Figure 3.38 contains the estimates where PSHGR and RHS are larger than the threshold from the estimating model in Equation 3.38. For the weighted scaled 2 estimates of $\sigma^2_{0k_1}$, PSHGR and RHS estimates have large differences for the simulation runs 13, 18 and 91. For the weighted unscaled estimates of σ^2_ϵ, the simulation run 97 produced large differences between PSHGR and RHS. For the weighted scaled 1 estimates of σ^2_ϵ, the simulation runs 26, 27, 42, 53, 54, 55, 81, 93, and 97 produced large differences between PSHGR and RHS. For the weighted scaled 2 estimates of σ^2_ϵ, the simulation runs 2, 5, 8, 10, 13, 18, 25, 41, 42, 46, 48, 49, 50, 54, 55, 59, 61, 63, 65, 68, 69, 73, 76, 78, 82, 92, and 98 produced large differences between PSHGR and RHS.

Figure ?? contains the estimates of $\sigma^2_{0k_1k_2}$ from the estimating model in Equation 3.39. This figure shows that simulation run 68 caused large differences between the PSHGR and RHS weighted unscaled, weighted scaled 1 and weighted scaled 2 estimates.

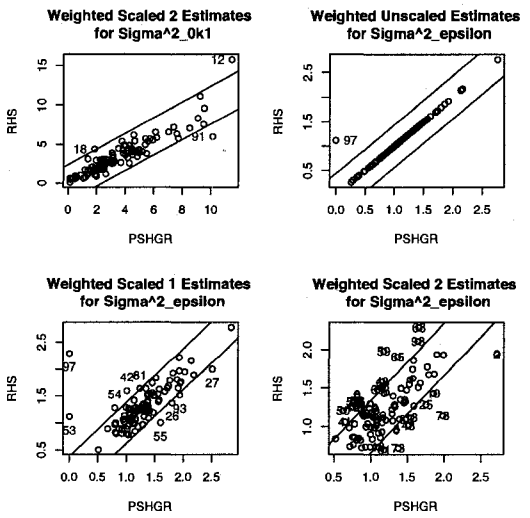

Figure 3.38: Comparison of PSHGR vs. RHS for Estimates from Equation 3.38

We next determine what we would expect the results to be for each of the estimating models. In Figure 3.11, the first row shows the estimates of the parameters when the bottom layer of clustering is removed. With this the variance of the $U_{0k_1k_2}$ term is put into the estimate of σ^2_ϵ, that becomes 1.5. There is some negative bias in the estimate of σ^2_{0k1}, due to the large intra-class correlation ($4/5.5 = 0.73$).

The second row shows the estimates of the parameters when the top layer of clustering is removed. The variance of $\sigma^2_{0k_1}$ should be put into the estimate of $\sigma^2_{0k_1k_2}$ to produce an estimate of 6. There is negative bias again, likely due to the large intra-class correlation ($6/6.5$=0.923).

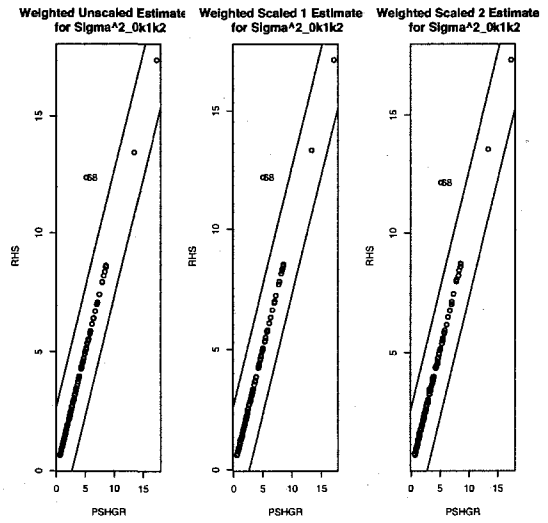

Figure 3.39: Comparison of PSHGR vs. RHS for Estimates from Equation 3.39

3.7.2 Negative Variance Components

There are situations in which the variance components are estimated to be negative. For example, in the case of a balanced random intercept model, say $y_{ik} = \beta_0 + \beta_{xk}x_k + \beta_{xij}x_{ij} + U_{0k} + \epsilon_{ik}$, the closed form estimate of the MLE of σ^2_{0k} in the unweighted case is $\frac{SSA}{k_S n_1} - \frac{\hat{\sigma}^2_\epsilon}{n_1}$, where k_S is the number of sampled clusters and n_1 is the number of elements sampled per cluster. In this case $SSA = \sum_k n_1[(\bar{y}_{\cdot k} - \bar{x}_{\cdot k}\beta) - (\bar{y}_{\cdot\cdot} - \bar{x}_{\cdot\cdot}\beta)]^2$, where the $\cdot$ in the subscript defines the variable being averaged over. There are cases when $\frac{\hat{\sigma}^2_\epsilon}{n_1}$ will be less than the term with SSA, resulting in a negative estimate for σ^2_{0k}. As is described in Searle et al. (1992) §3.7, when this occurs, the MLE of σ^2_{0k} becomes zero, and the estimate for σ^2_ϵ is adjusted. In this case, the $E(\hat{\sigma}^2_{0k}) = (1-p)E(\hat{\sigma}^2_{0k}|\hat{\sigma}^2_{0k} \geq 0)$ where p is the probability that $\hat{\sigma}^2_{0k}$ is negative. The density for the conditional distribution is not tractible, making the expected value difficult to obtain but can be estimated empirically in the simulations in this chapter. From Searle et al. (1992), p can be computed as $p = \Pr(\mathcal{F}^{K(N_1-1)}_{K-1} > (1-1/K)(1+n\frac{\sigma^2_{0k}}{\sigma^2_\epsilon}))$, where $\mathcal{F}^{K(N_1-1)}_{K-1}$ is a random variable with an F distribution with $K(N_1-1)$ and $K-1$ degrees of freedom.

This situation of negative estimated variance components occurs in some simulations, and will be noted as necessary. The simulations in this chapter are not balanced, and so the adjustment to the estimate of σ^2_ϵ is not computed. However Searle et al. (1992) show that the estimate of σ^2_ϵ without the adjustment (which are computed in the simulations) form an upper bound on the MLE.

3.7.3 RHS Sensitivity to the Number of Quadrature Points

To further investigate the differences between RHS and PSHGR, some of the simulation runs where the methods produce different estimates are examined. Specifically, various points from Figure 3.24 were run for a range of iteration points for RHS in the `gllamm()` function. The first point examined is in the first panel of Figure 3.24, an unweighted estimate of σ^2_{0k} from simulation run 2. The PSHGR and RHS results from a number of iteration points ranging from 15 to 30 are in Table 3.9. The table shows that the $\hat{\beta}_0, \hat{\beta}_2$ and $\hat{\sigma}^2_\epsilon$ are mostly unaffected by the iteration points . The estimates of $\hat{\sigma}^2_{0k}$ are quite sensitive, ranging from 9.32 to 17.83. Note from the log likelihood values, the maximum occurs at the parameter estimates from PSHGR. There are a number of iteration points that provide RHS estimates similar to the PSHGR estimates. Note that for 20 iteration points, the method did not converge. When the simulations were run for simulation set 11, the first converged simulation starting with 15 iteration points was chosen. Note that increasing the number of iteration points does not produce a monotonic increase in the log likelihood, as the lowest log likelihood occurred with 21 iteration points.

Method (Number of Iteration Points)	$\hat{\beta}_0$	$\hat{\beta}_2$	$\hat{\sigma}^2_\epsilon$	$\hat{\sigma}^2_{0k}$	Log Likelihood
PSHGR	3.24	-10.02	0.53	17.85	-475.04
RHS (15)	3.24	-10.02	0.53	11.68	-475.56
RHS (16)	3.25	-10.02	0.53	17.81	-475.04
RHS (17)	3.24	-10.02	0.53	17.83	-475.04
RHS (18)	3.23	-10.04	0.53	15.09	-475.15
RHS (19)	3.24	-10.02	0.53	17.83	-475.04
RHS (20)	NA	NA	NA	NA	NA
RHS (21)	3.24	-10.03	0.53	9.32	-476.39
RHS (22)	3.24	-10.02	0.52	15.06	-475.12
RHS (23)	3.24	-10.02	0.52	15.05	-475.12
RHS (24)	3.24	-10.02	0.53	17.85	-475.04
RHS (25)	3.24	-10.02	0.53	17.84	-475.04
RHS (26)	3.24	-10.02	0.53	17.85	-475.04
RHS (27)	3.25	-10.02	0.53	17.85	-475.04
RHS (28)	3.24	-10.02	0.53	17.84	-475.04
RHS (29)	3.24	-10.02	0.53	17.84	-475.04
RHS (30)	3.24	-10.02	0.53	17.85	-475.04

Table 3.9: Differences between RHS and PSHGR Estimated Parameters for Unweighted Estimates from Simulation Run 2 from Simulation Set 11, Estimating Model from Equation 3.34

The next point examined is in the first panel of Figure 3.24, an unweighted estimate of σ^2_{0k} from simulation run 53. The PSHGR and RHS results from a number of iteration points ranging from 15 to 30 are in Table 3.10. The table shows that the $\hat{\beta}_2$ and σ^2_ϵ are mostly unaffected by the number of iteration points. The estimates of $\hat{\beta}_0$ do vary between 2.28 and 2.83. The $\hat{\sigma}^2_{0k}$ are quite sensitive, ranging from 23.30 to 28.54. Note from the log likelihood values, the maximum occurs at the parameter estimates from PSHGR. There are a number of iteration points that provide RHS estimates similar to the PSHGR estimates. When the simulations were run for simulation set 11, the first converged simulation starting with 15 iteration points was chosen. Note that increasing the number of iteration points does not produce a monotonic increase in the log likelihood, as the lowest log likelihood occurred with 22 iteration points.

Method (Number of Iteration Points)	$\hat{\beta}_0$	$\hat{\beta}_2$	$\hat{\sigma}^2_\epsilon$	$\hat{\sigma}^2_{0k}$	Log Likelihood
PSHGR	2.62	-10.02	0.51	28.40	-470.85210
RHS (15)	2.28	-10.02	0.51	23.29	-470.98088
RHS (16)	2.61	-10.02	0.51	28.39	-470.85214
RHS (17)	2.71	-10.02	0.51	25.99	-470.87701
RHS (18)	2.61	-10.02	0.51	28.47	-470.85211
RHS (19)	2.65	-10.02	0.51	28.31	-470.85234
RHS (20)	2.62	-10.02	0.51	28.39	-470.85210
RHS (21)	2.62	-10.02	0.51	28.32	-470.85214
RHS (22)	2.60	-10.04	0.52	22.30	-471.06292
RHS (23)	2.83	-10.02	0.51	26.20	-470.87748
RHS (24)	2.62	-10.02	0.51	28.40	-470.85210
RHS (25)	2.61	-10.02	0.51	28.43	-470.85213
RHS (26)	2.62	-10.02	0.51	28.40	-470.85210
RHS (27)	2.61	-10.02	0.51	28.44	-470.85212
RHS (28)	2.61	-10.02	0.51	28.44	-470.85212
RHS (29)	2.59	-10.02	0.51	28.54	-470.85231
RHS (30)	2.62	-10.02	0.51	28.36	-470.85211

Table 3.10: Differences between RHS, and PSHGR Estimated Parameters for Unweighted Estimates from Simulation Run 53 from Simulation Set 11, Estimating Model from Equation 3.34

The next point examined is in the second panel of Figure 3.24, an weighted unscaled estimate of σ^2_{0k} from simulation run 53. The PSHGR and RHS results from a number of iteration points ranging from 15 to 30 are in Table 3.11. The table shows that the $\hat{\beta}_2$ and σ^2_ϵ are mostly unaffected by the number of iteration points. The estimates of $\hat{\beta}_0$ do vary between 1.89 and 2.32. The $\hat{\sigma}^2_{0k}$ are quite sensitive, ranging from 20.44 to 31.73. Note that there are no log likelihood values for PSHGR as there is no weighted likelihood. However, from the log likelihood values, the maximum occurs for PSHGR at iteration points 18, 26 and 30. Those corresponding estimates are close to the PSHGR estimates. When the simulations were run for simulation set 11, the first converged simulation starting with 15 iteration points was chosen. Note that increasing the number of iteration points does not produce a monotonic increase in the log likelihood, as the lowest log likelihood occurred with 25 iteration points.

Method (Number of Iteration Points)	$\hat{\beta}_0$	$\hat{\beta}_2$	$\hat{\sigma}^2_\epsilon$	$\hat{\sigma}^2_{0k}$	Log Likelihood
PSHGR	2.20	-10.13	0.52	31.67	NA
RHS (15)	1.89	-10.16	0.48	23.46	-1182.6300
RHS (16)	1.98	-10.17	0.49	23.53	-1182.2675
RHS (17)	2.24	-10.17	0.49	31.57	-1181.9494
RHS (18)	2.22	-10.17	0.48	31.68	-1181.9491
RHS (19)	NA	NA	NA	NA	NA
RHS (20)	2.31	-10.17	0.49	30.59	-1181.9550
RHS (21)	2.17	-10.17	0.49	31.00	-1181.9514
RHS (22)	2.07	-10.17	0.49	23.00	-1182.3078
RHS (23)	2.23	-10.17	0.49	26.95	-1182.0403
RHS (24)	2.32	-10.17	0.49	31.42	-1181.9517
RHS (25)	2.18	-10.16	0.50	20.44	-1182.7365
RHS (26)	2.22	-10.17	0.49	31.73	-1181.9491
RHS (27)	2.30	-10.17	0.49	30.78	-1181.9535
RHS (28)	2.14	-10.17	0.49	31.38	-1181.9504
RHS (29)	NA	NA	NA	NA	NA
RHS (30)	2.22	-10.17	0.49	31.73	-1181.9491

Table 3.11: Differences between RHS, and PSHGR Estimated Parameters for Weighted Unscaled Estimates from Simulation Run 53 from Simulation Set 11, Estimating Model from Equation 3.34

3.7.4 Description of the MSE Results

For the misspecification of fixed effects simulation set 1, the estimating model in Equation 3.3 both PSHGR and RHS prefered the weighted unscaled. This is surprising, because the estimating is the correct model with no informative sampling. The unweighted estimates do not have a smaller variance than the weighted estimates in this simulation. Also the differences between the $RRMSE$'s are very small, see Section 3.7.5. For example, the largest PSHGR $RRMSE$ is 0.0785 and the smallest is 0.0733. For the estimated model in Equation 3.4 the PSHGR and RHS estimates have different weighting schemes representing the lowest $RRMSE$. The RHS methodology has the lowest $RRMSE$ for the weighted unscaled estimates. As seen in Figures 3.2 and 3.12, there are some differences between the RHS and PSHGR weighted unscaled estimates of σ^2_{0k}. This is causing the mean of the RHS method to be lower than the mean of the PSHGR method, resulting different weighting schemes producing the lowest $RRMSE$. When the estimating model is from Equation 3.5, the estimation of the σ^2_ϵ is dominating the $RRMSE$ calculation. Because the weighted unscaled estimates are the smallest (i.e. closest to the true value of 0.5), both methodologies produce the smallest $RRMSE$ for the wieghted unscaled estimates. The $ARRMSE$ of PSHGR and RHS for estimated models in Equations 3.4 and 3.5 both prefer the unweighted estimates because of the smaller variances.

For the misspecification of fixed effects simulation set 4, for all the estimated models the PSHGR and RHS methods have the lowest $RRMSE$ with the unweighted estimates. Note the smaller variance from the unweighted estimators and that the weighting schemes are better at compensating for the informative sampling in the β_0, σ^2_{0k} and σ^2_ϵ parameters.

Likely, the reason why the unweighted estimates produce the smallest $RRMSE$ is because in the σ^2_{0k} and σ^2_ϵ estimates, the model misspecification in Equations 3.8 and 3.9 increase the bias and the unweighted estimates are the smallest. When the model misspecification is taken into account with the $ARRMSE$, the estimated model in Equation 3.8 has smallest $ARRMSE$ with the weighted scaled 1 estimates. However for $ARRMSE$ in from the estimated model in Equation 3.9, the compensation for the bias using the weighted estimates does not overcome the smaller variance of the unweighted estimates.

For the misspecification of the random effects simulation set 5, both estimated models from Equations 3.11 and 3.12 prefer the unweighted estimates. There is no informative sampling in this simulation set and the unweighted estimates have small variances. For the estimated model in Equation 3.12, only the $ARRMSE$ is computed, as the true value of the σ^2_{0k} parameter is zero.

For the misspecification of the random effects simulation set 6, the estimated model in Equation 3.14 produces the smallest $RRMSE$ with the weighted scaled 2 estimates. In this case, the unweighted estimates are not chosen because of both bias due to the informative sampling in the β_1 and σ^2_{1k} parameters. When determining which weighting scheme produces the lowest $RRMSE$, the σ^2_{1k} parameter dominates, and the weighted unscaled 2 estimates produce the lowest $RRMSE$.

For the misspecification of the random effects simulation set 7, the unweighted estimates produce the smallest $RRMSE$ (or $ARRMSE$) all the estimated models. This is because of the smaller variance of the unweighted estimates, the lack of informative sampling, and the small variance of σ^2_{2k}.

For the misspecification of the random effects simulation set 8, the estimating model in Equation 3.20 produced the smallest $RRMSE$ for PSHGR and RHS with the weighted scaled 1 estimates. The informative sampling produces bias in the unweighted estimates of β_2 and σ^2_{2k}. All the weighted schemes performed well with similar $RRMSE$. The $RRMSE$ for the PSHGR weighted estimates ranged from 0.2556 to 0.2204. The estimating model in Equation 3.21 produced the smallest $ARRMSE$ for PSHGR and RHS with the unweighted estimates. The largest contributers to the $ARRMSE$ are the estimates of σ^2_{0k}, and the unweighted estimates have the smallest values. The small variance on the β_0 and β_1 unweighted estimates also contribute to the smaller $ARRMSE$.

For the misspecification of the stratification layering simulation set 9, the $RRMSE$ (and RAsqMSE) is lowest for the unweighted estimates for all of the estimating models. For the estimating model in Equation 3.23, the true value of $\sigma^2_{01k.01k}$ is zero, and the estimates from this term were not included in the MSE calculations. The terms with the largest bias are the σ^2_ϵ estimates, of which the unweighted and weighted scaled 1 estimates produce the smallest $RRMSE$. The weighted scaled 1 estimates will for RHS produce a large $RRMSE$ for the σ^2_{02k} estimates (as explained about Figure 3.20). The unweighted estimates have slightly smaller variances, causing them to have the smallest $RRMSE$s. For the estimating models in Equations 3.24 and 3.26, the uweighted estimates produce the smallest $ARRMSE$s due to the smaller variances and the smaller bias on the σ^2_ϵ estimates.

For the misspecification of the stratification layering simulation set 10, the estimating model in Equation 3.28 the smallest $RRMSE$ is with the unwieghted estimates due to the low bias and variance of the estimates. For the estimating models in Equation 3.29 and

3.31 the $RRMSE$ is the smallest with the weighted unscaled estimates. This is because the model misspecification produces large positive bias on the σ^2_ϵ parameter and the weighted unscaled estimates have the smallest value. For the estimated model in Equation 3.29 the unweighted estimates produced the smallest $ARRMSE$ due to the smaller variances. For the estimated model in Equation 3.31, PSHGR and RHS produced different results. Notice that the PSHGR weighted scaled 1 estimates of σ^2_ϵ have a low 0.025 quantile, as seen in Figure 3.9, 3.24 and 3.26. The weighted scaled 1 estimates produced the lowest $ARRMSE$ for the RHS method and the weighted scaled 2 estimates produced the lowest $ARRMSE$ for the PSHGR method.

For the misspecification of the clustering layers, simulation set 11, the estimated model in Equation 3.33 contains no model misspecification. As expected, the $RRMSE$ for PSHGR is lowest for the unweighted estimates due to the minimal bias and smaller variance. However, the $RRMSE$ for RHS is lowest for the the weighted scaled 1 estimates. This is due to the very large bias in the unweighted estimate of σ^2_{02k}. The RHS weighted scaled 1 estimate of σ^2_{0k} is better behaved and generally has a smaller variance than the weighted scaled 2 estimates. For the estimated model in Equation 3.34, the $ARRMSE$ for both PSHGR and RHS favor the unweighted estimates due to the lower variance and the lack of informative sampling bias. For the estimating model in Equation 3.35, the $RRMSE$ favors the weighted unscaled estimates, as the $RRMSE$ is dominated by the σ^2_ϵ term and the weighted unscaled estimates are closest to the true value. When adjusting it for the anticipated values, the $ARRMSE$ for both RHS and PSHGR favor the unweighted estimates due ot the low variance and the lack of model misspecification bias. Finally, for

the estimated model in Equation 3.36, the $ARRMSE$ favors the unweighted estimates due to the smaller variance and the lack of informative sampling bias.

For the misspecification of the clustering layering simulation set 12, both estimating models contain model misspecification. For the estimated model in Equation 3.38, the $RRMSE$ is dominated by the bias in the σ^2_ϵ estimates and the weighted unscaled estimates have the lowest mean. For the RAsqMSE, the unweighted estimates produce the lowest numbers because of the low variance and minimal bias. For the estimated model in Equation 3.39, the $RRMSE$ is domiated by the bias in the σ^1_{0k1k2} estimates. The weighted scaled 1 estimates have the lowest $RRMSE$ for the σ^2_{0k10k2} parameter, so they also produce the lowest $RRMSE$ for the estimated model.

Tables 3.12 and 3.13 contain the numeric values of the $RRMSE$ and $ARRMSE$ for each simulation.

	Eqn. Num.	Weighting Scheme with Lowest MSE							
		Unweighted		Weighted Unscaled		Weighted Scaled 1		Weighted Scaled 2	
		RRMSE		*RRMSE*		*RRMSE*		*RRMSE*	
		P	R	P	R	P	R	P	R
Mis Fix 1	3.3	7.8594e-2	7.8587e-2	**7.3379e-2**	**7.3410e-2**	7.77096e-2	7.7317e-2	7.7096e-2	7.7317e-2
	3.4	3.2644e+4	2.8374e+4	3.1237e+4	**2.7002e+4**	**3.1143e+4**	3.2019e+4	3.1218e+4	2.9339e+4
	3.5	4.0222e+4	3.9975e+4	**3.1438e+4**	**3.1609e+4**	4.5315e+4	3.9533e+4	3.5561e+4	3.5635e+4
Mis Fix 4	3.7	**1.1561e-1**	**1.1791e-1**	2.0433e-1	2.0578e-1	2.0171e-1	2.0479e-1	1.8831e-1	1.9138e-1
	3.8	**1.4261e+4**	**1.3857e+4**	2.5140e+4	2.3402e+4	2.5084e+4	2.5006e+4	2.5119e+4	2.3461e+4
	3.9	**1.6398e+4**	**1.6261e+4**	2.6131e+4	2.6022e+4	1.8482e+4	1.8546e+4	2.3537e+4	2.3488e+4
Mis Ran 5	3.11	**9.4797e-2**	**9.4884e-2**	2.4933e-1	2.5020e-1	2.3252e-1	2.3242e-1	2.2132e-1	2.1994e-1
	3.12	—	—	—	—	—	—	—	—
Mis Ran 6	3.14	2.5834e-1	2.5846e-1	2.3064e-1	2.3085e-1	2.1135e-1	2.1013e-1	**2.0382e-1**	**2.0223e-1**
	3.15	—	—	—	—	—	—	—	—
Mis Ran 7	3.17	**9.5798e-2**	**9.7300e-2**	2.2984e-1	2.3196e-1	2.0004e-1	2.0086e-1	2.0581e-1	2.0591e-1
	3.18	—	—	—	—	—	—	—	—
Mis Ran 8	3.20	3.1733e-1	3.1731e-1	2.5557e-1	2.5613e-1	**2.2042e-1**	**2.2037e-1**	2.2716e-1	2.2711e-1
	3.21	—	—	—	—	—	—	—	—
Mis Strat 9	3.23	**6.8850e-1**	**6.8652e-1**	9.1607e-1	9.5862e-1	8.8713e-1	3.8200e+5	8.9394e-1	8.8134e-1
	3.24	—	—	—	—	—	—	—	—
	3.26	—	—	—	—	—	—	—	—
Mis Strat 10	3.28	**5.7122e-1**	**5.6701e-1**	7.3152e-1	7.3170e-1	6.9078e-1	6.901e-1	7.0105e-1	6.9815e-1
	3.29	—	—	—	—	—	—	—	—
	3.31	—	—	—	—	—	—	—	—
Mis Strat 11	3.33	**7.3889e-1**	1.1001e+4	8.4443e-1	4.5244e+4	8.2915e-1	**1.7074**	8.2344e-1	3.4923e+6
	3.34	—	—	—	—	—	—	—	—
	3.35	2.6340e+3	2.6309e+3	**2.0145e+3**	**2.0547e+3**	2.4009e+3	2.5901e+3	2.2253e+3	2.0565e+10
	3.36	—	—	—	—	—	—	—	—
Mis Clust 12	3.38	3.6977	3.6990	**2.4296**	**2.4310**	3.7148	3.6647	2.8694	2.9982
	3.39	1.7618e+1	1.7670e+1	**1.4982e+1**	**1.4745e+1**	1.5355e+1	1.5391e+1	1.5167e+1	1.4876+1

Table 3.12: Relative Root Mean Square Error ($RRMSE$) for each Simulation Set

		Eqn. Num.	Weighting Scheme with Lowest MSE: Unweighted *ARRMSE* P	Unweighted *ARRMSE* R	Weighted Unscaled *ARRMSE* P	Weighted Unscaled *ARRMSE* R	Weighted Scaled 1 *ARRMSE* P	Weighted Scaled 1 *ARRMSE* R	Weighted Scaled 2 *ARRMSE* P	Weighted Scaled 2 *ARRMSE* R
Mis	Fix 1	3.3	...	...	...	...	...	...	...	...
		3.4	**8.4129e-2**	**1.0717e-1**	2.7580e-1	2.8260e-1	2.4200e-1	2.4138e-1	2.5024e-1	2.5996e-1
		3.5	**1.6410e+1**	**1.4602e+1**	3.5597e+3	3.7144e+3	1.9264e+2	5.7353e+1	2.0935e+3	2.1438e+3
Mis	Fix 4	3.7	...	...	...	...	...	...	...	...
		3.8	6.7070e-1	6.7879e-1	3.7168e-1	3.8074e-1	**3.5823e-1**	**3.5876e-1**	3.6103e-1	3.6468e-1
		3.9	**8.1960e+1**	**7.8062e-1**	4.6589e+3	4.6490e+3	4.6319e+2	4.6718e+2	3.2840e+3	3.4033e+3
Mis	Ran 5	3.11	...	...	...	...	...	...	...	...
		3.12	**5.8642e-1**	**5.9620e-1**	1.3494	1.3550	1.3171	1.3100	1.3242	1.3283
Mis	Ran 6	3.14	...	...	...	...	...	...	...	...
		3.15	**7.6845e-1**	**7.6157e-1**	1.4761	1.3475	1.4331	1.4335	1.4456	1.4437
Mis	Ran 7	3.17	...	...	...	...	...	...	...	...
		3.18	**6.8822e-1**	**7.1810e-1**	2.1469	2.1804	1.4199	1.4362	1.7865	1.8101
Mis	Ran 8	3.20	...	...	...	...	...	...	...	...
		3.21	**7.5426e-1**	**7.6435e-1**	1.0430	1.0620	9.5318e-1	9.6936e-1	8.9989e-1	9.2430e-1
Mis	Strat 9	3.23	...	...	...	...	...	...	...	...
		3.24	**4.1869e-1**	**4.1724e-1**	2.7596	2.8129	2.7142	2.7195	2.7312	2.7314
		3.26	**1.6187**	**1.6190**	1.7263	1.7411	1.6772	1.6801	1.6952	1.7001
Mis	Strat 10	3.28	...	...	...	...	...	...	...	...
		3.29	**1.3675**	**1.3722**	2.6450	2.6218	2.7750	2.5768	2.5962	2.6028
		3.31	3.1449	3.1353	2.7032	2.7108	2.8265	**2.6567**	**2.6964**	2.7327
Mis	Strat 11	3.33	...	...	...	...	...	...	...	...
		3.34	**1.1087**	**1.0060**	1.2856	1.1450	1.2356	1.2328	1.2602	1.1857
		3.35	**1.2517**	**1.4311**	7.1681	1.3923e+1	2.6557	2.1000	5.7717	2.0565e+10
		3.36	**1.6316**	**1.6278**	4.3153	4.3419	4.3551	4.3528	4.3217	4.3244
Mis	Clust 12	3.38	**3.8637e-1**	**3.8616e-1**	5.2995e-1	5.1567e-1	4.7571e-1	4.4115e-1	4.5929e-1	4.5142e-1
		3.39	3.6999	3.7008	**2.4401**	**2.4438**	3.7290	3.6796	2.8812	3.0110

Table 3.13: Anticipated Relative Root Mean Square Error (*ARRMSE*) for each Simulation Set

3.7.5 Tables of True and Anticipated Parameter Values

For the computation of the $ARRMSE$ values, the anticipated parameter values are needed. The derivation of these values is in Section 3.7.1. They are also included in Tables 3.14 and 3.15 for reference.

		True (Anticipated) Parameter Values						
	Eqn. Num.	β_0	β_1	β_2	σ^2_{0k}	σ^2_ϵ	σ^2_{1k}	σ^2_{2k}
Mis Fix 1	3.3	1	-2	2	0.2	0.5		
	3.4	1(-5)	-2 (NA)	2 (2)	0.2 (36.2)	0.5 (0.5)		
	3.5	1(3)	-2 (-2)	2 (NA)	0.2 (0.2)	0.5 (100.5)		
Mis Fix 4	3.7	1	-2	2	0.2	0.5		
	3.8	1(-5)	-2 (NA)	2 (2)	0.2 (36.2)	0.5 (0.5)		
	3.9	1(3)	-2 (-2)	2 (NA)	0.2 (0.2)	0.5 (100.5)		
Mis Ran 5	3.11	1	-2	2		0.5	1	
	3.12	1(1)	-2 (-2)	2 (2)	0 (18)	0.5 (0.5)	1(NA)	
Mis Ran 6	3.14	1	-2	2		0.5	1	
	3.15	1(1)	-2 (-2)	2 (2)	0 (18)	0.5 (0.5)	1(NA)	
Mis Ran 7	3.17	1	-2	2		0.5		0.8
	3.18	1(1)	-2 (-2)	2 (2)	0 (1.6)	0.5 (16.5)		0.8(NA)
Mis Ran 8	3.20	1	-2	2		0.5		0.8
	3.21	1(1)	-2 (-2)	2 (2)	0 (1.6)	0.5 (16.5)		0.8(NA)

Table 3.14: True and Anticipated Parameter Values for Simulation Sets 1-8.

		True (Anticipated) Parameter Values							
	Eqn. Num.	β_0	β_1	σ^2_{01k}	σ^2_{02k}	$\sigma^2_{01k.02k}$ [13]	σ^2_ϵ	σ^2_{0k} or $\sigma^2_{0k_1}$	$\sigma^2_{0k_1k_2}$ or β_2
Mis Strat 9	3.23	-3	8	1	5	0	0.5		
	3.24	-3(1)	8 (NA)	1 (NA)	1 (NA)	0(NA)	0.5 (0.5)	0 (16)	
	3.26	-3(1)	8 (NA)	1 (NA)	1 (NA)	0 (NA)	0.5 (0.5)	0 (16)	
Mis Strat 10	3.28	-3	8				0.5	5	
	3.29	-3 (1)	8 (NA)				0.5 (16.5)	5 (5)	
	3.31	-3 (1)	8 (NA)				0.5 (16.5)	5 (5)	
Mis Strat 11	3.33	7	-8	1	5	0	0.5		-10
	3.34	7 (3)	-8 (NA)	1 (NA)	5 (NA)	0 (NA)	0.5 (NA)	0 (0)	-10 (-10)
	3.35	7 (2)	-8 (-8)	1 (1)	5 (5)	0 (0)	0.5 (NA)		-10 (NA)
	3.36	7 (-1)	-8 (NA)	1 (NA)	5 (NA)	0 (NA)	0.5 (25.5)	0 (16.5)	-10 (NA)
Mis Clust 12	3.38	5 (5)					0.5 (1.5)	5 (5)	1 (NA)
	3.39	5 (5)					0.5 (0.5)	5 (NA)	1 (6)

Table 3.15: True and Anticipated Parameter Values for Simulation Sets 9-12.

3.7.6 Computer Code

The simulations for PSHGR method were run using c-code I developed. This code may be found at `http://stat.cmu.edu` under the **Recent PhD Theses** link. The c-code uses the VMR library, downloaded from `http://www.stat.cmu.edu/~hseltman/`. It is in the Computer Programming, C/C++ section. The code uses `blas` functions, downloadable from `http://www.netlib.org`. The compilation instructions are commented in the beginning of the code. Along with the code are sample input files and the corresponding output file.

The simulations for the RHS method were run in `stata` using the `gllamm()` routine. The `gllamm()` routine can be found at `http://www.gllamm.org`.

Chapter 4

Sampling Weights in a GoM Model

The previous two chapters focus on the theory and practice of inserting weights reflecting the sampling design into linear mixed-effect models. This chapter focuses on incorporating the sampling design into other model-based analyses, specifically a Bayesian Grade of Membership (GoM) model. LME models contain variance structures that match many sampling designs. The natural variance component in the GoM model variance structure is a set of within-person random effects (latent variables) reflecting the general propensity of each survey respondent to respond positively to questions in a self-report. This needs to be adjusted to incorporate stratification and/or clustering effects.

To incorporate the sampling design into the GoM model, the standard Dirichlet prior on the GoM scores is replaced with a polytomous logistic mixed-effects regression prior. This prior can incorporate many different sampling designs, and is comparable to the LME models studied in Chapters 2 and 3. Guidelines for inserting weights into the joint distribution are developed. These weighting methods are derived from the PML methods of

Chapter 2, but the weighting of terms in the joint distribution (likelihood times prior) also depends on the parameters being estimated by that specific term. We call this *weighting based on the estimated parameter.* A simulation study is performed to analyze the effect of the weights on the GoM model.

Section 4.1 describes the standard unweighted GoM model and its derivation as seen in Erosheva (2002). Section 4.2 describes the changing to the polytomous logistic mixed-effects regression prior, first deriving the unweighted model. Weighting of the GoM model is discussed, and weighting based on the estimated parameter is derived. Section 4.3 describes some rotational indeterminancies in the GoM model, along with two known techniques for solving these indeterminancies. Sections 4.4 and 4.5 describe the details and descriptions of the simulation study. Section 4.6 summarizes the chapter. Section 4.7 collects together appendices providing further detail on this work. In particular section 4.7.3 provides a description and web-links for computer code used to conduct the simulations.

The contributions in this chapter involve both the GoM model analysis and incorporation of sampling weights. For the GoM model analysis, the polytomous logistic mixed-effects regression prior allows for model-based incorporation of the sampling design. It also provides a framework for the GoM model to be analyzed with longitudinal data (either with or without weights). With respect to sampling weights, we developed a principled way to incorporate sampling weights into a Bayesian model-based analysis, called weighting based on the estimated parameter. The simulation study provides a contribution regarding the actual performance of the weighting of the GoM model with the new prior. These simulations demonstrate the following: 1) When λ is fixed in the simulations, the mean of the

posterior distributions is generally similar to the simulations in which λ is unconstrained with an informative prior. However, the simulations in which λ is unconstrained with an informative prior have larger variance. This is true for the unweighted simulations, and mostly true for the weighted simulations. 2) The differences between the unweighted and weighted estimates of parameters of the polytomous logistic mixed-effects regression parameters behave similarly to the analogous differences seen in the Chapter 3 simulations, with a few exception noted in the simulation descriptions. Finally, 3) the estimates of the λ parameters appear robust to the sampling design and the type of weighting used in the estimation.

4.1 GoM Models

GoM models are a type of hierarchical Bayesian mixed-membership model used to analyze a variety of data, including depression-related psychiatric disorders, (Woodbury and Manton, 1989), the number of likely topics published in the *Proceedings of the National Academy of Sciences* in 1997-2000, and the number of underlying latent class disability profiles in the National Long Term Care Survey (NLTCS), (Airoldi et al., 2005). Erosheva (2002) provides an in-depth study of GoM models and disability data, including connections between GoM models and item response theory, factor analysis and principal component analysis.

4.1.1 Unweighted Derivation of the GoM Model

Following Erosheva (2002), the GoM model is comprised of extreme profiles and their conditional response probabilities. Let the data consist of J binary questions for I individuals.

Let $y_i = (y_{i1}, y_{i2}, ..., y_{iJ})$ be a vector of 0's and 1's representing the response of individual i on all J questions, $i = 1, ..., N$. A vector of GoM scores (latent variables) for each individual, $g_i = (g_{i1}, g_{i2}, ..., g_{iC})$, represents the mixture proportion of individual i in each of C unobservable latent classes. These GoM scores are non-negative and sum to 1 for each individual,

$$\sum_{c=1}^{C} g_{ic} = 1, \quad i = 1, ..., N. \tag{4.1}$$

Sole membership in a given class defines the pure response probability, λ_{cj}, for each of the J items of interest,

$$\lambda_{cj} = P(y_{ij} = 1 | g_{ic} = 1). \tag{4.2}$$

The following assumptions are made for the GoM model;

Assumption 1: The conditional probability of response of individual i to question j, given the GoM scores, is $P(y_{ij} = 1|g_i) = \sum_{c=1}^{C} g_{ic}\lambda_{cj}$.

Assumption 2: Conditional on the GoM scores, the responses y_{ij} are independent for different values of j, $(y_{ij_1} \perp y_{ij_2})|_{g_i}$.

Assumption 3: The responses y_{ij} are independent for different values of i, or $y_{i_1 j} \perp y_{i_2 j}$.

Assumption 4: The GoM scores, g_i, are realizations of a random vector with a Dirichlet distribution.

The GoM model in Erosheva (2002) allows the responses to the J items to be polytomous. For simplification, this thesis presents dichotomous response data only.

The GoM model is defined as

$$\begin{aligned} y_{ij}|g_i &\sim \text{Bernoulli}\left(\sum_{c=1}^{C} g_{ic}\lambda_{cj}\right) \\ g_i &\sim \text{Dirichlet}(\alpha_0\xi) \\ \lambda_{cj} &\sim \text{Beta}(\eta_{1cj}, \eta_{2cj}) \\ \alpha_0 &\sim \text{Gamma}(\tau_1, \tau_2) \\ \xi &\sim \text{Dirichlet}(\zeta), \end{aligned}$$

which contain a number of prior parameters and hyperparameters. Let η_{1cj} and η_{2cj} be parameters for the pure response probabilities, λ_{cj}. The prior parameters for the GoM scores are α_0, the prior sample size, and ξ, the prior proportions of the population elements in the underlying latent classes. The τ_1, τ_2 and ζ hyperparameters are set to be non-informative.

Erosheva (2002) augmented this model with latent variables, m_{ijc} to assign individual i to class c for question j, to ease the computations in the Bayesian MCMC estimation. Erosheva's fundamental representation theorem proves the equivalence between the GoM

model and the data augmented GoM model below,

$$\begin{aligned}
y_{ij}|m_{ijc},\lambda &\sim \text{Bernoulli}\left(\prod_{c=1}^{C}\lambda_{cj}^{m_{ijc}}\right)\\
m_{ijc}|g_i &\sim \text{Multinomial}(1, g_{i1}, ..., g_{iC})\\
g_i|\alpha_0,\xi &\sim \text{Dirichlet}(\alpha_0\xi)\\
\lambda_{cj} &\sim \text{Beta}(\eta_{1cj},\eta_{2cj})\\
\alpha_0 &\sim \text{Gamma}(\tau_1,\tau_2)\\
\xi &\sim \text{Dirichlet}(\zeta).
\end{aligned}$$

To solve for these parameters using a Bayesian MCMC algorithm, the joint distribution is computed as

$$\begin{aligned}
p(y,m,g,\lambda,\alpha_0,\xi) &\propto p(\lambda,\alpha)p(y,m,g|\lambda,\alpha_0,\xi)\\
&= p(\lambda)p(\alpha_0,\xi)\prod_{i=1}^{N}p(y_i,m_i,g_i|\lambda_i,\alpha_0,\xi)\\
&= p(\lambda)p(\alpha_0)p(\xi)\prod_{i=1}^{N}p(y_i|m_i,\lambda)p(m_i|g_i)p(g_i|\alpha_0,\xi).
\end{aligned}$$

This formulation assumes that $y_i|_{(m_i,\lambda)} \perp (g_i,\alpha)$ and $g_i|_{m_i} \perp (\lambda,\alpha)$ and $g_i|_{\alpha} \perp \lambda$.

Inserting the distributional assumptions into the joint distribution provides,

$$\begin{aligned} p(y,m,g,\lambda,\alpha) &\propto p(\lambda)p(\alpha)\prod_{i=1}^{N}[p(m_i|g_i)p(y_i|m_i\lambda)p(g_i|\alpha)] \\ &= p(\lambda)p(\alpha)\left(\prod_{i=1}^{N}\left[\frac{\Gamma(\sum_{c=1}^{C}\alpha_c)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\cdots\Gamma(\alpha_C)}g_{i1}^{\alpha_1-1}g_{i2}^{\alpha_2-1}\cdots g_{ik}^{\alpha_C-1}\right]\right) \\ &\times \prod_{i=1}^{N}\left(\prod_{j=1}^{J}\prod_{c=1}^{C}(g_{ic}\lambda_{cj}^{y_{ij}}(1-\lambda_{cj})^{1-y_{ij}})^{m_{ick}}\right) \end{aligned}$$

The complete conditionals are obtained for m_i, λ_{cj} and g_i,

$$\begin{aligned} m_i|- &\sim \text{Multinomial}(1,p_1,...,p_C) \;\; p_c \propto (g_{ic}\lambda_{cj}^{y_{ij}}(1-\lambda_{cj})^{(1-y_{ij})}) \\ \lambda_{cj}|- &\sim \text{Beta}(1+\sum_{i=1}^{N}y_{ij}m_{ijc}, 1+\sum_{i=1}^{N}(m_{ijc}-m_{ijc}y_{ij})) \\ g_i|- &\sim \text{Dirichlet}(\alpha_1+\sum_{j=1}^{J}m_{ij1}, \alpha_2+\sum_{j=1}^{J}m_{ij2}, \cdots, \alpha_C+\sum_{j=1}^{J}m_{ijC}). \end{aligned}$$

For α_0 and ξ, a Metropolis-Hastings step needs to be used. First consider α_0,

$$\begin{aligned} p(\alpha_0|-) &\propto p(\alpha_0)\prod_{i=1}^{N}\left[\frac{\Gamma(\alpha_0)}{\Gamma(\xi_1\alpha_0)\cdots\Gamma(\xi_C\alpha_0)}\prod_{c=1}^{C}g_{ic}^{\xi_c\alpha_0}\right] \\ &= \alpha_0^{\tau_1-1}e^{-\tau_2\alpha_0}\prod_{i=N}^{I}\left[\frac{\Gamma(\alpha_0)}{\Gamma(\xi_1\alpha_0)\cdots\Gamma(\xi_C\alpha_0)}\prod_{c=1}^{C}g_{ic}^{\xi_c\alpha_0}\right] \\ &= \alpha_0^{\tau_1-1}\exp\{-\alpha_0(\tau_2-\sum_{c=1}^{C}\sum_{i=1}^{N}\xi_k\log g_{ic}\}\left[\frac{\Gamma(\alpha_0)}{\Gamma(\xi_1\alpha_0)\cdots\Gamma(\xi_C\alpha_0)}\right] \end{aligned}$$

For the Metropolis-Hastings step, first draw a proposal point, α_0^* from the jumping distribution, and then calculate the proposal ratio. In this case, we draw a candidate point from a Gamma proposal distribution with parameters $\alpha = \gamma, \beta = \frac{\gamma}{\alpha_0^{(m)}}$, where $\alpha_0^{(r)}$ was

the last accepted value for α_0. The candidate point is accepted as the next element in the sample with probability $\min\{1, r_{\alpha_0}\}$. The proposal ratio, r_{α_0}, is

$$r_{\alpha_0} = \frac{p(\alpha_0^*|-)p(\alpha_0^{(r)}|\alpha_0^*)}{p(\alpha_0^{(r)}|-)p(\alpha_0^*|\alpha_0^{(r)})}.$$

Breaking this into two terms,

$$\begin{aligned}
r_{\alpha_0}(H) &= \frac{p(\alpha_0^{(r)}|\alpha_0^*)}{p(\alpha_0^*|\alpha_0^{(r)})} \\
&= \frac{\Gamma(\gamma, \gamma/\alpha_0^*)(\alpha_0^{(r)})}{\Gamma(\gamma, \gamma/\alpha_0^{(m)})(\alpha_0^*)} \\
&= \left(\frac{\alpha_0^{(r)}}{\alpha_0^*}\right)^{\gamma} \left(\frac{\alpha_0^{(r)}}{\alpha_0^*}\right)^{\gamma-1} \exp\left\{-\gamma\left(\frac{\alpha_0^{(r)}}{\alpha_0^*} - \frac{\alpha_0^*}{\alpha_0^{(r)}}\right)\right\} \\
r_{\alpha_0}(M) &= \frac{p(\alpha_0^*|-)}{p(\alpha_0^{(r)}|-)} \\
&= \left(\frac{\alpha_0^*}{\alpha_0^{(r)}}\right)^{\tau_1-1} \exp\left\{-(\alpha_o^* - \alpha_0^{(r)})(\tau_2 - \sum_{c=1}^{C}\sum_{i=1}^{N} \xi_c \log g_{ic}\right\} \\
&\times \left(\frac{\Gamma(\alpha_0^*)\prod_{c=1}^{C}\Gamma(\xi_c\alpha_0^{(r)})}{\Gamma(\alpha_0^{(r)})\prod_{c=1}^{C}\Gamma(\xi_c\alpha_0^*)}\right)^N.
\end{aligned}$$

Similarly, the Metropolis-Hastings step for ξ is derived,

$$\begin{aligned}
p(\xi|-) &\propto p(\xi)\prod_{i=1}^{N}\left[\frac{\Gamma(\alpha_0)}{\prod_{c=1}^{C}\Gamma(\xi_c\alpha_0)}\prod_{c=1}^{C} g_{ic}^{(\alpha_c-1)}\right] \\
&= \left[\frac{\Gamma(\alpha_0)}{\prod_{c=1}^{C}\Gamma(\xi_c\alpha_0)}\right]^N \exp\left\{\sum_{c=1}^{C}\sum_{i=1}^{N}(\alpha_0\xi_c - 1)\log g_{ic}\right\}. \qquad (4.3)
\end{aligned}$$

For the ξ_c's, a candidate point is drawn from a a Dirichlet proposal distribution centered at the previous sample value, $\text{Dirichlet}_{(\delta C\xi_1^{(r)},\cdots,\delta C\xi_c^{(r)})}(\xi^*)$. The candidate point is accepted

as the next element in the sample with probability $\min\{1, r_\xi\}$. The proposal ratio is r_ξ,

$$r_\xi = \frac{p(\xi^*|-)p(\xi^{(r)}|\xi^*)}{p(\xi^{(r)}|-)p(\xi^*|\xi^{(r)})}.$$

Breaking this into two terms,

$$\begin{aligned} r_\xi(M) &= \frac{p(\xi^*|-)}{p(\xi^{(r)}|-)} \\ &= \left(\frac{\prod_{c=1}^C \Gamma(\alpha_0\xi_c^{(r)})}{\prod_{c=1}^C \Gamma(\alpha_0\xi_c^*)}\right)^N \exp\left\{\sum_{c=1}^C \sum_{i=1}^N \alpha_0(\xi_c^* - \xi_c^{(r)}) \log g_{ic}\right\} \\ r_\xi(H) &= \frac{p(\xi^{(r)}|\xi^*)}{p(\xi^*|\xi^{(r)})} \\ &= \left(\frac{\prod_{c=1}^C \Gamma(\delta C\xi_c^{(r)})}{\prod_{c=1}^C \Gamma(\delta C\xi_c^*)}\right)\left(\frac{\prod_{c=1}^C \xi_c^{(r)(\delta C\xi_c^*-1)}}{\prod_{c=1}^C \xi_c^{*(\delta C\xi_c^{(r)}-1)}}\right). \end{aligned}$$

A sample from the posterior is obtained using MCMC with the complete conditionals and the Metropolis-Hastings steps.

4.2 Incorporation of the Sampling Design in the GoM Model

4.2.1 Polytomous Logistic Regression Prior in the GoM Model

Assuming clustering in the sampling design, all individuals in the population are not independent and the Assumption 3 from Section 4.1.1 no longer holds. Recall Assumption 1, that $P(y_{ij} = 1|g_i) = \sum_c g_{ic}\lambda_{cj}$. This suggests that the dependency between y_{i_1j} and y_{i_2j} is a result of the dependencies between the GoM scores, g_i's and/or the pure response prob-

abilities, λ_{cj}. Given that the GoM scores represent individual characteristics and the pure response probabilities represent class characteristics, I will represent dependencies between individuals through dependencies in their GoM scores. As seen in Chapters 2 and 3, linear mixed-effects models can model the dependencies of many sampling designs. Given that the GoM scores for an individual are positive and sum to 1, I propose using a polytomous logistic random-effects regression to model the effect of the sampling design on the GoM scores. Let y_{kij} represent the response of subject i in cluster k on question j. Similar changes in subscript are made on other variables. The assumptions from the original GoM score are now:

Assumption 1: The conditional probability of response of individual i in cluster k to question j, given the GoM scores, is $P(y_{kij} = 1|g_{ki}) = \sum_{c=1}^{C} g_{kic}\lambda_{cj}$.

Assumption 2: Conditional on the GoM scores, the responses y_{kij} are independent for different values of j, $(y_{kij_1} \perp y_{kij_2})|_{g_{ki}}$.

Assumption 3: The responses of y_{kij} for all subjects i in the same cluster k are dependent.

Assumption 4: The GoM scores, g_{ki} are realizations of a random vector with a polytomous logistic random-effects distribution.

These assumptions allow GoM model analysis on data from a survey.

The updated GoM model is defined as

$$
\begin{aligned}
y_{kij}|m_{kijc},\lambda &\sim \text{Bernoulli}\left(\prod_c \lambda_{cj}^{m_{kijc}}\right) \\
m_{kijc}|\psi_i &\sim \text{Multinomial}(1, g_{ki1}, ..., g_{kiC}) \\
& \quad g_{kic} = \frac{\exp\{\psi_{kic}\}}{\sum_{c=1}^{C}\exp\{\psi_{kic}\}} \\
\psi_{ic}|X,Z,U,\beta &\sim N(X_i\beta_c + Z_iU_c, \sigma_\psi^2), c = 1, \cdots, C-1 \qquad (4.4) \\
& \quad \psi_{kiC} = 0, \psi_{kic} = \log\left(\frac{g_{kic}}{g_{kiC}}\right), c = 1, ..., C-1 \\
\lambda_{cj} &\sim \text{Beta}(\eta_{1cj}, \eta_{2cj}) \\
\beta_c &\sim \text{Normal}(\mu_\beta, \Sigma_\beta) \\
U_c &\sim \text{Normal}(0, \Omega) \qquad (4.5) \\
\sigma_\psi^2 &\sim \text{Scaled Inv } \chi^2(\nu, s_\psi^2)
\end{aligned}
$$

This utilizes the LME framework from Chapter 2 to insert the effect of the sampling design on the GoM model. While the subscript on y_{kij} denotes a clustered only design, more complex designs change the above model trivially, by changing the structure of the X and Z matrices to incorporate the stratification and clustering information as was done in Chapters 2 and 3.

Similar to the unweighted GoM model, this model is estimated using MCMC. Before considering sampling weights, we consider estimation of this unweighted model. The joint

conditional distribution is

$$\begin{aligned} p(y,m,\psi,\lambda,\beta,U,\sigma^2_\psi,X,Z) &\propto p(\beta,U,\sigma^2_\psi)p(\lambda)p(y,m,\psi|\lambda,\beta,U,\sigma^2_\psi,X,Z) \qquad (4.6)\\ &= p(\beta)p(U)p(\sigma^2_\psi)p(\lambda)p(y|m,\lambda)p(m|\psi)p(\psi|\beta,U,\sigma^2_\psi,X,Z) \end{aligned}$$

In the last equation, we assume that $(y|m,\lambda) \perp (\beta,U,\sigma^2_\psi,\psi)$ and $(m|\psi) \perp (\beta,U,\sigma^2_\psi)$ and that $(\psi|\beta,U,\sigma^2_\psi) \perp \lambda$. Continuing,

$$\begin{aligned} p(y,m,\psi,\lambda,\beta,U,\sigma^2_\psi,X,Z) &\propto p(\beta)p(U)p(\sigma^2_\psi)p(\lambda)\\ &\times \left[\prod_{k=1}^{K}\prod_{i=1}^{N_k}\left(\prod_{j=1}^{J}\prod_{c=1}^{C} p(y_{kij}|m_{kijc},\lambda_{cj})\right)\right]\left[\prod_{k=1}^{K}\prod_{i=1}^{N_k}\prod_{j=1}^{J}\prod_{c=1}^{C} p(m_{kijc}|\psi_{kic})\right]\\ &\times \prod_{c=1}^{C-1}\prod_{k=1}^{K}\prod_{i=1}^{N_k} p(\psi_{kic}|\beta,U,\sigma^2_\psi) \end{aligned}$$

Recall that k (of K) indexes clusters, i (of N_k) indexes individuals in clusters, j (of J) indexes questions, c (of C) indexes latent classes. Writing $p(U)\prod_k p(\psi_k|\beta,U,\sigma^2_\psi) = \prod_k(\prod_i\prod_c p(\psi_{kic}|\beta,U,\sigma^2_\psi))p(U_k)$ will be useful for the insertion of sampling weights in the

next section. Inserting in the distributional forms provides

$$\begin{aligned} p(y, m, \psi, \lambda, \beta, U, \sigma^2_\psi, X, Z) \quad \propto \quad & \exp\left\{-\frac{1}{2}\sum_{c=1}^{C}(\beta_c - \mu_\beta)^T\Sigma_\beta^{-1}(\beta_c - \mu_\beta)\right\} \\ & \times \quad (\sigma^2_\psi)^{-(\frac{\nu}{2}+1)}\exp\left\{-\frac{\nu s^2_\psi}{2\sigma^2_\psi}\right\}\left[\prod_{c=1}^{C}\prod_{j=1}^{J}\lambda_{cj}^{\eta_{1cj}-1}(1-\lambda_{cj})^{\eta_{2cj}-1}\right] \\ & \times \quad \prod_{k=1}^{K}\prod_{i=1}^{N_k}\prod_{j=1}^{J}\prod_{c=1}^{C}\left[\frac{\exp(\psi_{kic})}{\sum_{c_1=1}^{C}\exp(\psi_{kic_1})}\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{1-y_{kij}}\right]^{m_{kijc}} \\ & \times \quad \prod_{k=1}^{K}\left[\prod_{i=1}^{N_k}\left(\prod_{c=1}^{C-1}(\sigma^2_\psi)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2_\psi}(\psi_{kic} - X_{ki}\beta_c - Z_{ki}U_c)^2\right\}\right)\right. \\ & \times \quad \left.\exp\left\{-\frac{1}{2}U_{kc}^T\mathcal{O}^{-1}U_{kc}\right\}\right] \end{aligned} \tag{4.7}$$

Note that the prior on U_{kc} uses the cluster version of the covariance matrix, $\mathcal{O}$, instead of the entire covariance of U, Ω. See Section 2.1.1 for more details. From this, we can get the complete conditionals for Gibbs steps in the MCMC. The complete conditionals for the

parameters associated with the polytomous logistic regression are

$$
\begin{aligned}
p(\beta_c|-) &\propto \exp\left\{-\frac{1}{2}(\beta_c-\mu_c)^T\Sigma_\beta^{-1}(\beta_c-\mu_c)\right\}\prod_{k=1}^{K}\prod_{i=1}^{N_k}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\} \\
&\sim \text{Normal}(\mu_1,\Sigma_1) \\
&\quad \mu_1=\left(\Sigma_\beta^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K}\sum_{i=1}^{N_k}X_{ki}^TX_{ki}\right)^{-1}\left(\Sigma_\beta^{-1}\mu_c+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K}\sum_{i=1}^{N_k}X_{ki}^T(\psi_{kic}-Z_{ki}U_c)\right) \\
&\quad \Sigma_1=\left(\Sigma_\beta^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K}\sum_{i=1}^{N_k}X_{ki}^TX_{ki}\right)^{-1} \\
p(U_c|-) &\propto \prod_{k=1}^{K}\prod_{i=1}^{N_k}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\}\exp\left\{-\frac{1}{2}U_{kc}^T\Omega^{-1}U_{kc}\right\} \\
&\sim \text{Normal}(\mu_2,\Sigma_2) \\
&\quad \mu_2=\left(\mathcal{O}^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K}\sum_{i=1}^{N_k}Z_{ki}^TZ_{ki}\right)^{-1}\left(\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K}\sum_{i=1}^{N_k}Z_{ki}^T(\psi_{kic}-X_{ki}\beta_c)\right) \\
&\quad \Sigma_2=\left(\mathcal{O}^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K}\sum_{i=1}^{N_k}Z_{ki}^TZ_{ki}\right)^{-1} \\
p(\sigma_\psi^2|-) &\propto (\sigma_\psi^2)^{-(\frac{\nu}{2}+1)}\exp\left\{-\frac{\nu s_\psi^2}{2\sigma_\psi^2}\right\}\prod_{k=1}^{K}\prod_{i=1}^{N_k}\prod_{c=1}^{C-1}(\sigma_\psi^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kc}-X_k\beta_c-Z_kU_{kc})^2\right\} \\
&\sim \text{Scaled Inv }\chi^2\left(N(C-1)+\nu,\frac{\left(\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{c=1}^{C-1}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_{kc})^2\right)+\nu s_\psi^2}{N(C-1)+\nu}\right)
\end{aligned}
$$

The complete conditionals associated with the pure response probabilities and data augmented values are

$$\begin{aligned}
p(\lambda_{cj}|-) &\propto \prod_{k=1}^{K}\prod_{i=1}^{N_k}\left[\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right]^{m_{kijc}}\lambda_{cj}^{\eta_{1cj}-1}(1-\lambda_{cj})^{\eta_{2cj}-1} \\
&\sim \text{Beta}\left(\sum_{k=1}^{K}\sum_{i=1}^{N_k}y_{kij}m_{kijc}+\eta_{1cj}, \sum_{k=1}^{K}\sum_{i=1}^{N_k}(1-y_{kij})m_{kijc}+\eta_{2cj}\right) \\
p(m_{kij}|-) &\propto \prod_{c=1}^{C}\left[\frac{\exp\{\psi_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right]^{m_{kijc}} \\
&\sim \text{Multinomial}\left(1, \frac{\exp\{\psi_{ki1}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{1j}^{y_{kij}}(1-\lambda_{1j})^{1-y_{kij}}, \cdots, \right. \\
&\quad \left. \frac{\exp\{\psi_{kiC}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{Cj}^{y_{kij}}(1-\lambda_{Cj})^{1-y_{kij}}\right)
\end{aligned}$$

Finally, a Metropolis step is needed for ψ,

$$\begin{aligned}
p(\psi_{kic}|-) &\propto \prod_{c=1}^{C}\prod_{j=1}^{J}\left[\frac{\exp\{\psi_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\right]^{m_{kijc}} \\
&\times \exp\left\{-\frac{1}{2\sigma_{\psi}^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\}
\end{aligned}$$

Let a candidate point be drawn from a Normally distribution, with the mean at the previous MCMC value and a variance of $\sigma^2_{\psi\text{jmp}}$. The candidate point is accepted as the next element

in the sample with probability $\min\{1, r_{\psi_{kic}}\}$, where

$$\begin{aligned} r_{\psi_{kic}} &= \frac{p(\psi^*_{kic}|-)}{p(\psi^{(r)}_{kic}|-)} \\ &= \prod_{c=1}^{C}\prod_{j=1}^{J}\left[\frac{\exp\{\psi^*_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi^*_{kic_1}\}}\frac{\sum_{c_1=1}^{C}\exp\{\psi^{(r)}_{kic_1}\}}{\exp\{\psi^{(r)}_{kic}\}}\right]^{m_{kijc}} \\ &\times \exp\left\{-\frac{1}{2\sigma^2_\psi}\left[(\psi^*_{kic} - X_{ki}\beta_c - Z_{ki}U_c)^2 - (\psi^{(r)}_{kic} - X_{ki}\beta_c - Z_{ki}U_c)^2\right]\right\} \\ &= \prod_{c=1}^{C}\left[\frac{g^*_{kic}}{g^{(r)}_{kic}}\right]^{\sum_{j=1}^{J} m_{kijc}} \exp\left\{-\frac{1}{2\sigma^2_\psi}\left[(\psi^*_{kic} - X_{ki}\beta_c - Z_{ki}U_c)^2 - (\psi^{(r)}_{kic} - X_{ki}\beta_c - Z_{ki}U_c)^2\right]\right\} \end{aligned}$$

where $g_{kic} = \frac{\exp\{\psi_{ic}\}}{\sum_{c_1=1}^{C}\psi_{kic_1}}$ where $\psi_{kiC} = 0$. These complete conditionals and Metropolis-Hastings steps can be implemented using MCMC algorithms.

4.2.2 Weighting the Logistic Regression GoM Model

Given that the GoM model has been modified to incorporate the sampling design, we next evaluate if the sampling weights provide any additional information. In Chapters 2 and 3, where the sampling design was also modeled, the sampling weights did help compensate for informative sampling but not model misspecification. We next investigate the effect of the sampling weights on informative sampling with the GoM model. The effect of model misspecification and sampling weights on the GoM model will be discussed in future work.

Recall from Section 2.2.2 that psuedo-maximum likelihood can estimate census likelihood equations with weighted sample likelihood equations. When using PML on the GoM model, similar issues arise as with the LME model, namely, when should weights be inserted and does inserting the weights in different areas affect the results? More specifically,

should we

1. Add weights to Equation 4.7 and have them propagate through to the complete conditionals and Metropolis-Hastings steps?

2. Add weights directly to the complete conditionals?

3. Use an alternate weighting method?

4. Scale of the weights as was done with the LME's (see Section 2.4.2)?

These issues are explored next.

Adding Sampling Weights to Equation 4.7

To add weights to the GoM model likelihood, consider the methods described in Chapter 2 and 3 for weighting LME mdoels. This PML estimation of the census joint distribution is similar to the RHS method from Section 2.2.3. The subscript w denotes this weighted joint distribution. This provides a weighted joint distribution of

$$\begin{aligned} p_w(y, m, \psi, \lambda, \beta, U, \sigma^2_\psi, X, Z) \quad \propto \quad & p(\beta)p(\sigma^2_\psi)p(\lambda) \\ & \times \prod_{k=1}^{K}\prod_{i=1}^{N_k}\left(\prod_{j=1}^{J}\prod_{c=1}^{C} p(y_{kij}|m_{kijc}, \lambda_{cj})\right)^{w_{ki}} \\ & \times \prod_{k=1}^{K}\prod_{i=1}^{N_k}\left[\prod_{j=1}^{J}\prod_{c=1}^{C} p(m_{kijc}|\psi_{kic})\right]^{w_{ki}} \\ & \times \prod_{c=1}^{C-1}\prod_{k=1}^{K}\left[\left(\prod_{i=1}^{N_k} p(\psi_{kic}|\beta, U, \sigma^2_\psi)^{w_{i|k}}\right) p(U_k)\right]^{w_k} \end{aligned}$$

The derivation of all the complete conditionals and Metropolis Hastings steps are in Section 4.7.1. An issue with this weighting is that weights propogate to the complete conditionals in unexpected ways.

Consider, the complete conditional for m_{kijc} is

$$\begin{aligned} p_w(m_{ki}|-) &\propto \prod_{j=1}^{J}\prod_{c=1}^{C}\left(g_{kic}\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right)^{m_{kijc}w_{ki}} \\ m_{ki}|- &\sim \text{Multinomial}(1,p_1,...,p_C) \;\; p_c \propto (g_{kic}\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})})^{w_{ki}}, \end{aligned}$$

where $g_{kic} = \frac{\exp\{\psi_{kic}\}}{\sum_{c_1}\exp\{\psi_{kic_1}\}}$.The m_{kij} parameter describes a characteristic of an individual as opposed to being a summary variable for the finite population (such as the λ_{cj}'s). It is not immediately clear that, for example, the probabilities in the multinomial distribution for m_{kij} should be raised to the power of the weight so that it can represent more people. If provided the complete conditionals based on the census, the weights do not seem to have a place in the complete conditionals of m_{kij}. Similar arguments hold for weights in the ψ_{kic} complete conditionals. This leads to the next option for incorporating the sampling weights.

Adding Sampling Weights to the Complete Conditionals

Consider estimating the census complete conditional with weighted sample complete conditionals. The subscript wCC below denotes the result from weighting the complete conditionals. The complete derivation of these complete conditionals and Metropolis-Hastings steps are in Section 4.7.2.

Consider the complete conditionals for λ and m with this weighting scenario,

$$
\begin{aligned}
p_{wCC}(\lambda_{cj}|-) &\propto \prod_{k=1}^{K_s}\prod_{i=1}^{n_k}\left[\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right]^{m_{kijc}w_{ki}}\lambda_{cj}^{\eta_{1cj}-1}(1-\lambda_{cj})^{\eta_{2cj}-1} \\
&\sim \text{Beta}\left(\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}y_{kij}m_{kijc}w_{ki}+\eta_{1cj}, \sum_{k=1}^{K_s}\sum_{i=1}^{n_k}(1-y_{kij})m_{kijc}w_{ki}+\eta_{2cj}\right) \\
p_{wCC}(m_{kij}|-) &\propto \prod_{c=1}^{C}\left[\frac{\exp\{\psi_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right]^{m_{kijc}} \\
&\sim \text{Multinomial}\left(1, \frac{\exp\{\psi_{ki1}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{1j}^{y_{kij}}(1-\lambda_{1j})^{1-y_{kij}}, \cdots,\right. \\
&\quad \left.\frac{\exp\{\psi_{kiC}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{Cj}^{y_{kij}}(1-\lambda_{Cj})^{1-y_{kij}}\right)
\end{aligned}
$$

With this wCC weighting, some components of the joint distribution are treated differently in different complete conditionals. For example, in $p_{wCC}(\lambda_{cj}|-)$, the $\left[\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{1-y_{kij}}\right]$ term from the posterior is weighted. However, the same term in $p_{wCC}(m_{kij}|-)$ is not weighted. Treating a component from the posterior differently in different complete conditionals appears unprincipled. This leads to the new weighting scheme below.

Weighting based on the Estimated Parameter

A more principled way to add sampling weights to the GoM model is to weight the term of the joint distribution based on the parameter upon which that term is used to make an inference. If the term of the joint distribution is making inferences only on individual parameters, then it does not need to be weighted. If the term of the joint distribution is making inferences on any group parameters (or parameters that more than one individual is dependent upon), then it should be weighted. Call this weighting based on the estimated

parameter, and subscript the estimates with wEP.

To understand the reasoning for this, first define two different types of distributions (or conditional distributions); 1) distributions providing information for at least one group parameter and 2) distributions providing information for individual/cluster parameter or priors with no estimable parameters. Consider the unweighted joint distribution as an example

$$\begin{aligned} p(y,m,\psi,\lambda,\beta,U,\sigma^2_\psi,X,Z) \quad &\propto \quad p(\beta)p(U)p(\sigma^2_\psi)p(\lambda) \\ &\times \quad \left[\prod_{k=1}^{K}\prod_{i=1}^{N_k}\left(\prod_{j=1}^{J}\prod_{c=1}^{C} p(y_{kij}|m_{kijc},\lambda_{cj})\right)\right] \\ &\times \quad \left[\prod_{k=1}^{K}\prod_{i=1}^{N_k}\prod_{j=1}^{J}\prod_{c=1}^{C} p(m_{kijc}|\psi_{kic})\right] \\ &\times \quad \prod_{c=1}^{C-1}\prod_{k=1}^{K}\prod_{i=1}^{N_k} p(\psi_{kic}|\beta,U,\sigma^2_\psi) \end{aligned} \tag{4.8}$$

The likelihood portion of the joint distribution has three components; 1) $p(y_{kij}|m_{kijc},\lambda_{cj})$, 2) $p(m_{kijc}|\psi_{kic})$ and 3)$p(\psi_{kic}|\beta,U,\sigma^2_\psi)$. Note that $p(y_{kij}|m_{kijc},\lambda_{cj})$ uses the data in y_{kij} to gain more information about both m_{kijc} and λ_{cj}. Here m_{kijc} is an individual parameter which applies only to individual ki. However, λ_{cj} is a group parameter, affecting more than just the ki^{th} individual. The $p(y_{kij}|m_{kijc},\lambda_{cj})$ terms combine information across many y_{kij} to estimate the group parameter λ_{cj}. Because of this, the $p(y_{kij}|m_{kijc},\lambda_{cj})$ term provides information for a group parameter. Similarily, $p(\psi_{kic}|\beta,U,\sigma^2_\psi)$ combines information across many ψ_{kic} to estimate the group parameters β,U and σ^2_ψ. Contrast this to $p(m_{kijc}|\psi_{kic})$. The ψ_{kic} parameter only pertains to the ki^{th} individual. The $\psi_{kic} =$

$X_{ki}\beta_c + Z_{ki}U_{kc} + \epsilon_{ki}$, so it is a function of group parameters. However, as noted just after Equation 4.6, the assumption is that $m_{kijc}|_{\psi_{kic}} \perp (\beta_c, U_{kc}, \sigma^2_\psi)$. Therefore, the distribution $p(m_{kijc}|\psi_{kic})$ provides information about the individual parameter ψ_{kic} but not any of the group parameters.

Next consider the prior distributions used to form the joint distribution. I will classify the priors into two different groups, 1) non data-scalable priors that will not be weighted and 2) data-scalable priors that will be weighted. Non data-scalable priors are priors that do not change dimension regardless of the size of the data. The $p(\beta), p(\sigma^2_\psi)$ and $p(\lambda)$ do not change dimension if the number of individuals or clusters increase. However, $p(U)$ is a data-scalable prior, as the dimensions of U change as the number of clusters changes. Recall that the dimension of U is $KQ \times 1$, where K is the number of clusters, as defined in Section 2.1.1. When a sample is taken, then the dimension of the prior is $K_sQ \times 1$ where K_s is the number of sampled clusters.

In this wEP weighting scheme, the distributions providing information for at least one group parameter and the data-scalable priors are weighted. The distributions providing information for individual/cluster parameters and the non data-scalable priors are not weighted.

The wEP weighted joint distribution becomes

$$\begin{aligned} p_{wEP}(y,m,\psi,\lambda,\beta,U,\sigma_\psi^2,X,Z) \propto\ & p(\beta)p(\sigma_\psi^2)p(\lambda) \\ & \times \left[\prod_{k=1}^{K}\prod_{i=1}^{N_k}\left(\prod_{j=1}^{J}\prod_{c=1}^{C} p(y_{kij}|m_{kijc},\lambda_{cj})\right)^{w_{ki}}\right] \\ & \times \prod_{k=1}^{K}\prod_{i=1}^{N_k}\left[\prod_{j=1}^{J}\prod_{c=1}^{C} p(m_{kijc}|\psi_{kic})\right] \\ & \times \prod_{c=1}^{C-1}\prod_{k=1}^{K}\left[\left(\prod_{i=1}^{N_k} p(\psi_{kic}|\beta,U,\sigma_\psi^2)^{w_{i|k}}\right) p(U_k|\Omega)\right]^{w_k} \end{aligned}$$

Inserting the distributional forms provides

$$\begin{aligned} p_{wEP}(y,m,\psi,\lambda,\beta,U,\sigma_\psi^2,X,Z) \propto\ & \exp\left\{-\frac{1}{2}\sum_{c=1}^{C}(\beta_c-\mu_\beta)^T\Sigma_\beta^{-1}(\beta_c-\mu_\beta)\right\} \\ & \times (\sigma_\psi^2)^{-(\frac{\nu}{2}+1)}\exp\left\{-\frac{\nu s_\psi^2}{2\sigma_\psi^2}\right\}\left[\prod_{c=1}^{C}\prod_{j=1}^{J}\lambda_{cj}^{\eta_{1cj}-1}(1-\lambda_{cj})^{\eta_{2cj}-1}\right] \\ & \times \prod_{k=1}^{K_s}\prod_{i=1}^{n_k}\prod_{j=1}^{J}\prod_{c=1}^{C}\left[\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{1-y_{kij}}\right]^{m_{kijc}w_{ki}} \\ & \times \prod_{k=1}^{K_s}\prod_{i=1}^{n_k}\prod_{j=1}^{J}\prod_{c=1}^{C}\left[\frac{\exp(\psi_{kic})}{\sum_{c_1=1}^{C}\exp(\psi_{kic_1})}\right]^{m_{kijc}} \\ & \times \prod_{k=1}^{K_s}\left[\prod_{i=1}^{n_k}\left(\prod_{c=1}^{C-1}(\sigma_\psi^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\}\right)^{w_{i|k}}\right. \\ & \left. \times \exp\left\{-\frac{1}{2}U_{kc}^T\mathcal{O}^{-1}U_{kc}\right\}\right]^{w_k} \end{aligned}$$

This leads to the following complete conditionals for the regression variables

$$
\begin{aligned}
p_{wEP}(\beta_c|-) &\propto \exp\left\{-\frac{1}{2}(\beta_c-\mu_c)^T\Sigma_\beta^{-1}(\beta_c-\mu_c)\right\}\prod_{k=1}^{K_s}\prod_{i=1}^{n_k}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\}^{w_{ki}} \\
&\sim \text{Normal}(\mu_1,\Sigma_1) \\
&\quad \mu_1=\left(\Sigma_\beta^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{Ks}\sum_{i=1}^{n_k}w_{ki}X_{ki}^TX_{ki}\right)^{-1}\left(\Sigma_\beta^{-1}\mu_c+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}X_{ki}^T(\psi_{kic}-Z_{ki}U_c)\right) \\
&\quad \Sigma_1=\left(\Sigma_\beta^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}X_{ki}^TX_{ki}\right)^{-1} \\
p_{wEP}(U_c|-) &\propto \prod_{k=1}^{K}\left[\prod_{i=1}^{n_k}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_{kc})^2\right\}^{w_{i|k}}\exp\left\{-\frac{1}{2}U_{kc}^T\mathcal{O}^{-1}U_{kc}\right\}\right]^{w_k} \\
&\sim \text{Normal}(\mu_2,\Sigma_2) \\
&\quad \mu_2=\left(\sum_{k=1}^{K}w_k\mathcal{O}^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}Z_{ki}^TZ_{ki}\right)^{-1}\left(\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}Z_{ki}^T(\psi_{kic}-X_{ki}\beta_c)\right) \\
&\quad \Sigma_2=\left(\sum_{k=1}^{K}w_k\mathcal{O}^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}Z_{ki}^TZ_{ki}\right)^{-1} \\
p_{wEP}(\sigma_\psi^2|-) &\propto (\sigma_\psi^2)^{-(\frac{\nu}{2}+1)}\exp\left\{-\frac{\nu s_\psi^2}{2\sigma_\psi^2}\right\}\prod_{k=1}^{K_w}\prod_{i=1}^{n_k}\prod_{c=1}^{C-1}(\sigma_\psi^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kc}-X_k\beta_c-Z_kU_c)^2\right\}^{w_{ki}} \\
&\sim \text{Scaled Inv }\chi^2\left(\nu_1,s_1^2\right) \\
&\quad \nu_1=\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}(C-1)+\nu \\
&\quad s_1^2=\frac{\left(\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}\sum_{c=1}^{C-1}w_{ki}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right)+\nu s_\psi^2}{\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}(C-1)+\nu}
\end{aligned}
$$

The complete conditionals for the augmented data and pure response probabilities are

$$
\begin{aligned}
p_{wEP}(\lambda_{cj}|-) &\propto \prod_{k=1}^{K_s}\prod_{i=1}^{n_k}\left[\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right]^{m_{kijc}w_{ki}}\lambda_{cj}^{\eta_{1cj}-1}(1-\lambda_{cj})^{\eta_{2cj}-1} \\
&\sim \text{Beta}\left(\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}y_{kij}m_{kijc}w_{ki}+\eta_{1cj},\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}(1-y_{kij})m_{kijc}w_{ki}+\eta_{2cj}\right) \\
p_{wEP}(m_{kij}|-) &\propto \prod_{c=1}^{C}\left[\frac{\exp\{\psi_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\left(\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right)^{w_{ki}}\right]^{m_{kijc}} \\
&\sim \text{Multinomial}\left(1,\frac{\exp\{\psi_{ki1}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\left[\lambda_{1j}^{y_{kij}}(1-\lambda_{1j})^{1-y_{kij}}\right]^{w_{ki}},\cdots,\right. \\
&\quad \left.\frac{\exp\{\psi_{kiC}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\left[\lambda_{Cj}^{y_{kij}}(1-\lambda_{Cj})^{1-y_{kij}}\right]^{w_{ki}}\right)
\end{aligned}
$$

Finally, the Metropolis-Hastings step for ψ is

$$
\begin{aligned}
p_{wEP}(\psi_{kic}|-) &\propto \prod_{c=1}^{C}\prod_{j=1}^{J}\left[\frac{\exp\{\psi_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\right]^{m_{kijc}} \\
&\times \exp\left\{-\frac{1}{2\sigma_{\psi}^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\}
\end{aligned}
$$

Let a candidate point be drawn from a Normal distribution, with the mean at the previous MCMC value and a variance of $\sigma^2_{\psi\text{jmp}}$. The candidate point is accepted as the next element

in the sample with probability $\min\{1, r_{\psi_{kic}}\}$, where

$$\begin{aligned} r_{(wEP)\psi_{kic}} &= \frac{p(\psi^*_{kic}|-)}{p(\psi^{(r)}_{kic}|-)} \\ &= \prod_{c=1}^{C}\prod_{j=1}^{J}\left[\frac{\exp\{\psi^*_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi^*_{kic_1}\}}\frac{\sum_{c_1=1}^{C}\exp\{\psi^{(r)}_{kic_1}\}}{\exp\{\psi^{(r)}_{kic}\}}\right]^{m_{kijc}} \\ &\times \exp\left\{-\frac{1}{2\sigma^2_\psi}\left[(\psi^*_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2-(\psi^{(r)}_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right]\right\} \\ &= \prod_{c=1}^{C}\left[\frac{g^*_{kic}}{g^{(r)}_{kic}}\right]^{\sum_{j=1}^{J}m_{kijc}} \exp\left\{-\frac{1}{2\sigma^2_\psi}\left[(\psi^*_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2-(\psi^{(r)}_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right]\right\} \end{aligned}$$

One concern regarding the wEP weighting involve the weights in the $p_{wEP}(m_{kij}|-)$ distribution. When the weights are large, then $\left(\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right)^{w_{ki}}$ may become very small, basically zero to machine precision. This will be addressed by scaling the weights.

Scaling of the Weights

In Chapters 2 and 3, the scaling of the weights played a role in the estimation of parameters, especially the variance components. The scaling in LME models was introduced to reduce the bias in the variance components. Recall that the scaled 1 weightings from Section 2.4.2 adjust the conditional weights, $w_{i|k}$, so $\sum_i w^{s1}_{i|k}$ equals the effective sample size for cluster k, as defined in Potthoff et al. (1992). The scaled 2 weights from Section 2.4.2 adjust the conditional weights so that $\sum_i w^{s2}_{i|k}$ equals the cluster sample size for cluster k, n_k. The scaling of the weights will also be used in the simulations in this chapter.

The posterior variances of the parameters are affected by the scaling of the weights. By

weighting the data, the sample size affectively becomes $\sum_k \sum_i w_{ki}$. When the weights are unscaled, this is an estimate of the size of the *population*, which is larger than the sample size and will create smaller posterior variances. For the simulations in this chapter, the weights are scaled analogous to the scaled 2 weights from Section 2.4.2 so that $\sum_{i=1}^{n_k} w^s_{i|k} = n_k$, where $w^s_{i|k}$ represents the scaled conditional weight from cluster k and n_k Unlike the LME models, it is not clear that the scaling of the cluster weights, w_k, will have no affect on the estimates. The cluster weights are scaled so that $\sum_{k=1}^{K_s} w^s_k = K_s$ where w^s_k represents the scaled cluster weight and K_s represents the number of sampled clusters. The effect of the scalings of the weights is an area for further investigation.

4.3 Indeterminancies in the GoM model

4.3.1 GoM model, Factor Analysis and Rotations

The GoM model and factor analysis models both contain latent class structures designed to find factors to explain interrelationships among observable variables. Woodbury and Manton (1989), Marini et al. (1996) and Erosheva (2002) compare and contrast the latent structures of the GoM models and the factor analytic models. Unfortunately, these models both have rotational indeterminancies.

Rotational indeterminancies in factor analysis are well known and researched; see any statistical multivariate analysis text such as Johnson and Wichern (1992). Rotational indeterminancies in the GoM model have not been previously documented. To see where these rotations are inserted in the GoM model, consider again Assumption 1 from Section

4.1.1 and Assumption 1 from Section 4.2.1,

$$P_{kij} = P(y_{kij} = 1|g_{ki}, \lambda_{cj}) = \sum_{c=1}^{C} g_{kic}\lambda_{cj} = g_{ki}^T\lambda_j$$

where $g_{ki} = (g_{ki1}, g_{ki2}, \cdots, g_{kiC})^T$ are the GoM scores for individual ki and the pure response probabilities for item j are $\lambda_j = (\lambda_{1j}, \lambda_{2j}, \cdots, \lambda_{Cj})^T$. Collecting all the P_{kij} into a matrix, see that

$$P = G\Lambda$$

where $G = (g_1^T, \cdots, g_N^T)^T$ and $\Lambda = (\lambda_1, \cdots, \lambda_J)$. If R is an invertible matrix subject to suitable restrictions discussed below, define,

$$\begin{aligned} G^* &= GR \\ \Lambda^* &= R^{-1}\Lambda \end{aligned}$$

Now G^* and R^* are a new rotation for G and R. The definition of the probabilities of response remains the same,

$$P = G^*\Lambda^*$$

Restrictions on the matrix R come from Equations 4.1 and 4.2. Namely,

$$\begin{aligned} GR1_{C\times 1} &= 1 \\ GR &\geq 0 \\ R^{-1}\Lambda &\geq 0 \\ R^{-1}\Lambda &\leq 1 \end{aligned}$$

where all inequalities are elementwise. The set of matrices R satisfying these conditions usually has positive Lebesgue measure in the space of invertible matrices R, hence finding these rotations in an MCMC algorithm is possible. Described next are two ways to work with the rotational indeterminancies in the GoM model, using informative priors and fixing $\lambda_{\cdot}$ parameters.

4.3.2 Informative Priors

In frequentist analysis, a number of specific rotations are defined for factor analysis. The purpose of these rotations is to find factor loadings that are easily interpretable. There are a variety of standard rotations that are used, such as orthogonal rotations (including varimax, quartimax and equimax) and oblique rotations (including promax).

In Bayesian factor analysis, using informative (or subjective) priors uniquely determines the rotation, see Kaufman and Press (1973) and Rowe (2001). The formulation of the GoM model in Sections 4.2.1, and 4.2.2 allows for informative priors, especially the prior for λ_{cj} and the specification η_{1cj} and η_{2cj}. Whenever informative priors are used in the remainder

of this chapter, it is clearly stated.

4.3.3 Fix λ Parameters

GoM models and item response theory (IRT) models contain similar latent class structures, Erosheva (2005). As an example, the National Assessment of Educational Progress (NAEP) uses IRT models to estimate proficiency scores (analogous to the GoM scores or g_i's) of students on different skills, see the special issue of the Journal of Educational Measurement (1992). In estimating the proficiency scores, the NAEP model first estimates item parameters (analogous to the pure response probabilities, λ's in the GoM model) ignoring the survey design completely, then assumes the item parameters are fixed and produces random draws of the proficiency scores accounting for the survey design, see von Davier et al. (2007). In fixing the λ's when estimating the g's, the NAEP estimation avoids the rotational indeterminancy and provides a precedent for this approach. An explanation of the estimation of the proficiency scores and item parameters in NAEP based upon Mislevy and Sheehan (1989b) follows.

Informative Stratified Sampling and GoM Models

Mislevy and Sheehan (1989a,b) argue that differential probabilities of sampling in a stratified sampling model do not affect estimation of item parameters in an IRT model, but may affect estimation of proficiency scores. Their argument is shown below in the context of the GoM model using the Dirichlet prior from Section 4.1.1. Similar results hold for the GoM model with logistic regression prior from Section 4.2.1.

Suppose that the GoM scores, g_i and the augmented data inclusion variables, m_{ijc} are observed. Consider the likelihood portion of Equation 4.3,

$$p(y, m, g|\lambda, \alpha_0, \xi) = \prod_{i=1}^{N} p(y_i, m_i, g_i|\lambda, \alpha_0, \xi).$$

Suppose the census data are stratified, where the distribution of the GoM scores differs in each stratum. Let h_i represent the stratum indicator for element i. Let $f(g_i|\alpha_{h_i}, \xi_{h_i}, h_i)$ be the distribution of the GoM scores in stratum h_i. Let π_s be the proportion of the *population* in stratum s and assume that

$$f(g|\alpha, \xi) = \sum_{s=1}^{H} \pi_s f(g|\alpha_s, \xi_s, h_i = s).$$

The likelihood becomes

$$\begin{aligned} p(y, m, g, H|\lambda, \alpha_0, \xi) &= \prod_{i=1}^{N} f(H = h_i|\lambda, \pi, \alpha, \xi) f(y_i, m_i, g_i|h_i, \lambda, \pi, \alpha, \xi) \\ &= \prod_{i=1}^{N} \Pr(H = h_i|\pi) f(y_i|m_i, \lambda) p(m_i|g_i) f(g_i|h_i, \xi_{h_i}, \alpha_{h_i}) \\ &= \prod_{s=1}^{H} \pi_s^{N_s} \times \prod_{s=1}^{H} f(y_s|m_s, \lambda) p(m_s|g_s) f(g_h|h_s, \xi_{h_s}, \alpha_{h_s}) \end{aligned}$$

where the y_s represent all responses from people in stratum s. There are corresponding definitions for m_s and g_s. Note that the π_s is distinct from the rest of the likelihood and consistency for these parameters is derived using standard results on the multinomial distribution, as the N_s grow. The second term is a product of H likelihoods with a common

λ parameter. Bradley and Gart (1962) show conditions for consistency when the likelihood is made up of separate populations that have distinct population parameters (such as g_s, m_s, ξ_{h_s} and α_{h_s}.) and a few common parameters, such as λ.

Suppose a sample is taken and that the proportion of sampled elements within a stratum does not equal the proportion of population elements in the stratum. Now define

$$f^*(g|\alpha,\xi) = \sum_{s=1}^{H} \pi_s^* f(g|\alpha_s, \xi_s, h_i = s),$$

where π_s^* is the *sample* proportion of the elements in stratum s. Then the likelihood becomes

$$\begin{aligned} p(y,m,g,H|\lambda,\alpha_0,\xi) &= f(H|\lambda,\pi,\alpha,\xi) f(y,m,g|H,\lambda,\pi,\alpha,\xi) \\ &= \prod_{i=1}^{n} \Pr(H = h_i|\pi^*) f(y_i|m_i,\lambda) p(m_i|g_i) f(g_i|h_i,\xi_{h_i},\alpha_{h_i}) \\ &= \prod_{s=1}^{H} \pi_s^{*N_s} \times \prod_{s=1}^{H} f(y_s|m_s,\lambda) p(m_s|g_s) f(g_s|h_s,\xi_{h_s},\alpha_{h_s}). \end{aligned}$$

Similar to the case where the census was taken, consistent estimates of of λ, α_h's, ξ_h's can be obtained, but we can not reconstruct $f(g|\alpha,\xi) = \sum_s \pi_s f(g|\alpha_s,\xi_s,h_i = s)$ because of our inability to estimate π, the population proportions when the sampling design uses biased π^*'s.

This argument is used in NAEP to show that the item response parameters (or λ's in the GoM model) are not affected by the sampling design, whereas the achievement scores (or g's in the GoM model) are affected by the sampling design. Next is a brief description

of the implementation of the Mislevy and Sheehan (1989b) results in NAEP.

National Assessment of Educational Progress

The National Assessment of Educational Progress (NAEP) is a national assessment program that regularly tests students in grades 4, 8 and 12 on a variety of academic subjects. NAEP provides an important operational example of the methodology advocated by Mislevy and Sheehan (1989a,b). The data from NAEP are analyzed and published in the National's Report Card (see `http://nces.ed.gov/nationsreportcard/`) for comparative analysis across years. There are many complexities in NAEP's design that are derived from the limited time to administer tests (often about one hour) while at the same time producing reliable and valid assessments. The goal of NAEP is to produce reliable estimates of proficiency for specific population subgroups in various academic subjects.

Group proficiencies are estimated using three stages as described in von Davier et al. (2007). In the *Scaling* stage, an IRT model is fit to the data to estimate the item response parameters (equivalent to the λ's in the GoM model). These IRT models do not account for the sampling design. After estimation of the item response parameters, they are considered fixed in the remaining analysis. The justification for this is from Mislevy and Sheehan (1989b) as stated in Thomas (2000). In the *Conditioning* stage, marginal maximum likelihood is used to estimate the mean and variance of the proficiency for students in the population given the students individual covariate values (i.e. the g's in the GoM model). Using that mean and variance, random draws (also called multiple imputations or plausible values) are obtained from the examinees posterior latent variable and are used to create

subgroup estimates. In the *Variance Estimation* stage, multiple imputation and jackknife approaches are used to estimate subgroup variances.

The simulation study in this chapter provides results of weighting on the GoM model when there are informative priors on the λ parameters, and when the λ's are fixed.

4.4 GoM Simulation Study Set-Up

The role of the sampling weights in the GoM model is analyzed using a simulation study. These simulations are designed for a number of comparisons; 1) the difference between the polytomous logistic mixed-effects prior parameters (GoM estimates) when λ is fixed versus when λ is unconstrained with an informative prior, 2) the difference between unweighted and wEP weighted estimates, both when λ is fixed and when λ is unconstrained with an informative prior, and 3) the difference between unweighted and wEP weighted estimates of λ.

The simulated sampling design is the same in all the simulations; two top-level strata, and within each stratum there are clusters. Elements from the clusters are then sampled. There are three different levels of informativeness,

Non-Informative (Non): Clusters and elements are sampled according to the size of an independently generated random variable.

Informative Clusters (Clust): Clusters with larger random effects are oversampled. Elements are sampled according to the size of an independently generated random variable.

Informative Individuals (Indiv): Clusters are sampled according to the size of an independently generated random variable. Elements with larger random errors, ϵ_{ik} are oversampled.

4.4.1 True Values in the Simulated Model

The simulated model contains $C = 2$ underlying classes and $J = 5$ questions. The *population* has 2 strata, with 20 clusters per stratum ($Q = 40$ population clusters) and 250 elements per cluster for a population size of $N = 10,000$. The polytomous regression prior has two X covariates which are indicators of stratum inclusion and no intercept. The population Z matrix is of dimension (10000 $\times$ 80) or ($N \times KQ$). The regression function for a given element is

$$\psi_{hki1} = -1I_{h==1} + 0.5I_{h==2} + U_{1k}I_{h==1} + U_{2k}I_{h==2} + \epsilon_{hki1}$$

$$U_{01k} \sim N(0, 0.04),\ U_{02k} \sim N(0, 0.64),\ \epsilon_{hkic} \sim N(0, 0.25)$$

where $h = 1$ or 2 denotes stratum inclusion and k represents the cluster. This LME is similar to the generating model from Equation 3.32 in Simulation Set 11 in Chapter 3 modified for only one level of stratification. This LME contains two random slopes, each on a stratum inclusion indicator variable. There is no data to estimate the values of U_{2k} when the element is in stratum 1. As such, the posterior of U_{2k} for values of k in stratum 1 matches the prior. Those values are not reported in the simulation. The reverse holds true for U_{1k} and stratum 2.

	class 1	class 2
Question 1	0.765	0.050
Question 2	0.723	0.407
Question 3	0.447	0.410
Question 4	0.642	0.483
Question 5	0.950	0.250

Table 4.1: True Value of Simulated λ

In this model, $C = 2$ so the ψ_{hkic} is only defined for $c = 1$, as the baseline class for the polytomous logistic regression is $c = 2$. In other words, this is a logistic regression as there are only two classes. The mean GoM score for someone in stratum 1 is $E(g_{1ki}) = (\frac{\exp\{-1\}}{1+\exp\{-1\}}, \frac{1}{1+\exp\{-1\}}) = (0.27, 0.73)$, and for stratum 2 is $E(g_{2ki}) = (\frac{\exp\{0.5\}}{1+\exp\{0.5\}}, \frac{1}{1+\exp\{0.5\}}) = (0.62, 0.38)$.

The true value of $\lambda_{cj} = P(y_{hkij} = 1|g_{ki} = c)$ for $C = 2$ classes and $J = 5$ questions is in Table 4.1. These correspond roughly to a "sick" class and a "healthy" class. Prior parameters are set at $\mu_\beta = 0I_{2\times 1}, \Sigma_\beta = 10I_{2\times 2}$. The value of Ω is

$$\Omega = \begin{bmatrix} 0.01 & 0.001 \\ 0.001 & 0.01 \end{bmatrix} \otimes I_{K_s \times K_s}$$

Throughout the simulations, the estimates of σ^2_ψ tended to drift. To control this, an informative prior was used, with prior degrees of freedom $\nu = 200$, and prior mean equalling the true value of $s^2_\psi = 0.25$. Changing this back to a non-informative prior is discussed in the future work. For the simulations below that use informative priors on λ, the values of η_{1cj}, η_{2cj} are listed in the descriptions below.

	UNnon FixL	UNclust FixL	Unindiv FixL	wEPnon FixL	wEPclust FixL	wEPindiv FixL
Acceptance Ratio of ψ's	45.5%	45.5%	45.5%	40.1%	42.3%	44.2%
Number of Iterations	200K	200K	200K	400K	300K	300K
Number of Burn-In Iterations	10K	10K	10K	10K	10K	10K
Amount of Thinning	20	20	20	20	20	20

Table 4.2: Notes for the MCMC simulation when λ is Fixed

	UNnon Uncon-strL	UNclust Uncon-strL	Unindiv Uncon-strL	wEPnon Uncon-strL	wEPclust Uncon-strL	wEPindiv Uncon-strL
Acceptance Ratio of ψ's	45.4%	45.5%	45.5%	40.8%	43.1%	44.4%
Value of $\eta_{1cj} + \eta_{2cj}$	100	100	100	200	300	150
Number of Iterations	300K	200K	200K	500K	500K	300K
Number of Burn-In Iterations	10K	10K	10K	10K	30K	10K
Amount of Thinning	20	20	20	20	20	20

Table 4.3: Notes for the Informative Prior (Unconstrained) λ MCMC simulation

4.4.2 MCMC Notes

Details of the twelve MCMC simulations presented in the graphs below are in Tables 4.2 and 4.3 and discussed next.

For the implementation of the MCMC, there is a Metropolis-Hastings step for the ψ_{hkic}. The acceptance ratio is computed for each ψ_{hkic} for the 1000 elements in the sample and the average acceptance ratio over the sampled elements is in Table 4.2 and 4.3. The target acceptance ratio is about 40%. The jumping distribution for the ψ is normal, centered at the previous ψ value with variance 1.

It is not possible to state exactly when a sample from the MCMC algorithm represents a random sample from the posterior distribution. Assessing convergence of an MCMC chain is delicate and I used a number of tools. The MCMC chains have a burn-in period to remove effects of initial values and are thinned (to remove iteration to iteration dependencies and

save computing resources). These values were chosen by running each simulations for an initial 5000 iterations (no burn-in and no thinning) and using the Raftery & Lewis convergence diagnostics in the `R` package `boa` (Raftery and Lewis, 1992). The Raftery and Lewis convergence diagnostics provides guidelines on MCMC burn-in, thinning and number of runs to achieve an estimate of the 0.025 and 0.975 percentiles with an accuracy of ± 0.005 with a probability of 95%. Convergence of the MCMC runs after the burn-in and thinning was monitored using the Heidelberger & Welch convergence diagnostics, also in `boa` (Heidelberger and Welch, 1983). The Raftery & Lewis and Heidelberger & Welch diagnostics are summarized in Cowles and Carlin (1996). The Raftery & Lewis total number of iterations, burn-in and thinning were altered based on the results of the Heidelberger & Welch diagnostics, if necessary. In addition, I visually examine the plots for any evidence of nonstationarity. Finally, I also examine the trace of the log likelihood using the same set of methods. For comparison, Erosheva (2002) analyzed a GoM model with 2 underlying classes and the Dirichlet prior, using 100,000 MCMC iterations with 10,000 iterations of burn-in and thinned by taking every 10^{th} iteration for the sample.

Most of the parameters passed convergence tests, however each of the simulations had non-convergence in some of the the random effects (i.e. the U_{01k} and U_{02k}). Most of the failed tests were the Heidelberger & Welch halfwidth tests, though two of the U's failed both the stationary test and the halfwidth test. However in mixed-effects regressions, the specific values of the U_{0k}'s are usually not of interest. It is the variance of the random effect that is of interest, and that value is reported in the simulations below. The variance of the random effects was computed for each MCMC iteration. The chain of variances of

Label	Meaning
UN	Unweighted analyses
wEP	Weighted analysis using wEP weights
non	Non-informative sampling
clust	Clusters with large random effects, U_{01k} or U_{02k}, are oversampled
indiv	Individuals with large random error, ϵ_{hki}, are oversampled
UnconstrL	Unconstrained λ's
FixedL	Fixed λ's

Table 4.4: Labels on GoM Simulations

the U's did converge for all of the simulations.

4.4.3 Presentation Format

The format of the presentation of the results is similar to that of Chapter 3, described in Section 3.3. The differences between the Chapter 3 format and the format below are described here.

For each parameter there are two vertical lines, one grey and one light blue. The grey line is the simulated parameter value. Because there is known shrinkage towards the prior mean in Bayesian analyses, the light blue line indicates the mean of the unweighted non-informative mean. Comparisons of the effects of informative sampling and weights are compared to the unweighted non-informative estimates. Each line is labeled above. The labels are in Table 4.4.

4.5 GoM Simulation Results

Five simulation plots are shown below to evaluate the comparisons noted above; 1) the difference between unweighted and wEP weighted polytomous logistic mixed-effects regres-

sion prior parameters when λ is fixed versus when λ is unconstrained with an informative prior, 2) the difference between unweighted and wEP weighted estimates, both when λ is fixed and when λ is unconstrained with an informative prior, and 3) the difference between unweighted and wEP weighted estimates of λ.

4.5.1 Unweighted Results of Sampling Design Parameters (GoM Scores)

The results from this simulation are in Figure 4.1. First, examine the difference in the estimates where λ is fixed versus where λ is unconstrained with an informative prior. The unweighted estimates under the non-informative sampling scheme when λ is fixed versus when λ is unconstrained are very similar. For the unweighted estimates under informative cluster sampling scheme, the unconstrained λ estimates have larger posterior spread for the β's. The posterior spread on the variance components are similar for the fixed versus the unconstrained λ estimates. For the unweighted estimates under the informative individual sampling scheme, the unconstrained λ estimates also have larger posterior spread for the β's. The posterior spread on the variance components are similar for the fixed versus the unconstrained λ estimates.

Next, examine the difference in the fixed-effects under the different sampling schemes. As mentioned earlier, the β estimates exhibit shrinkage towards their prior mean of zero. When the sampling design oversamples clusters with large random effects (U_{01k} or U_{02k}) then the estimates of the β's increase, as expected. Similarly, when the sampling design oversamples individuals with large random errors (ϵ_{hkic}) the estimates of β increase.

Finally, examine the difference in the variance components under the different sampling

schemes. The estimates of σ^2_ψ are consistent across sampling schemes. I expect that the random error variance would decrease when the individuals are informative sampled based on ϵ_{hkic}. I believe that the underestimation of σ^2_ψ is not seen in these simulations because of the informative prior placed on the parameter, as discussed in Section 4.4.1. The loosening of this informative prior is discussed in the future work. The estimates of $\text{Var}(U_k)$ for Stratum 1 and Stratum 2 are smaller for the informative cluster sampling than the non-informative sampling, as expected. Under the informative individual sampling scheme, the estimates of $\text{Var}(U_k)$ for Stratum 1 and Stratum 2 are larger than the informative cluster sampling scheme, as expected. There appears to be some positive bias for the estimates of $\text{Var}(U_k)$ Stratum 1 and some negative bias for the estimates of $\text{Var}(U_k)$ Stratum 2 when compared to the noninformative sampling scheme. This bias is not well understood, however it may be due to the lack of bias on the σ^2_ψ estimate.

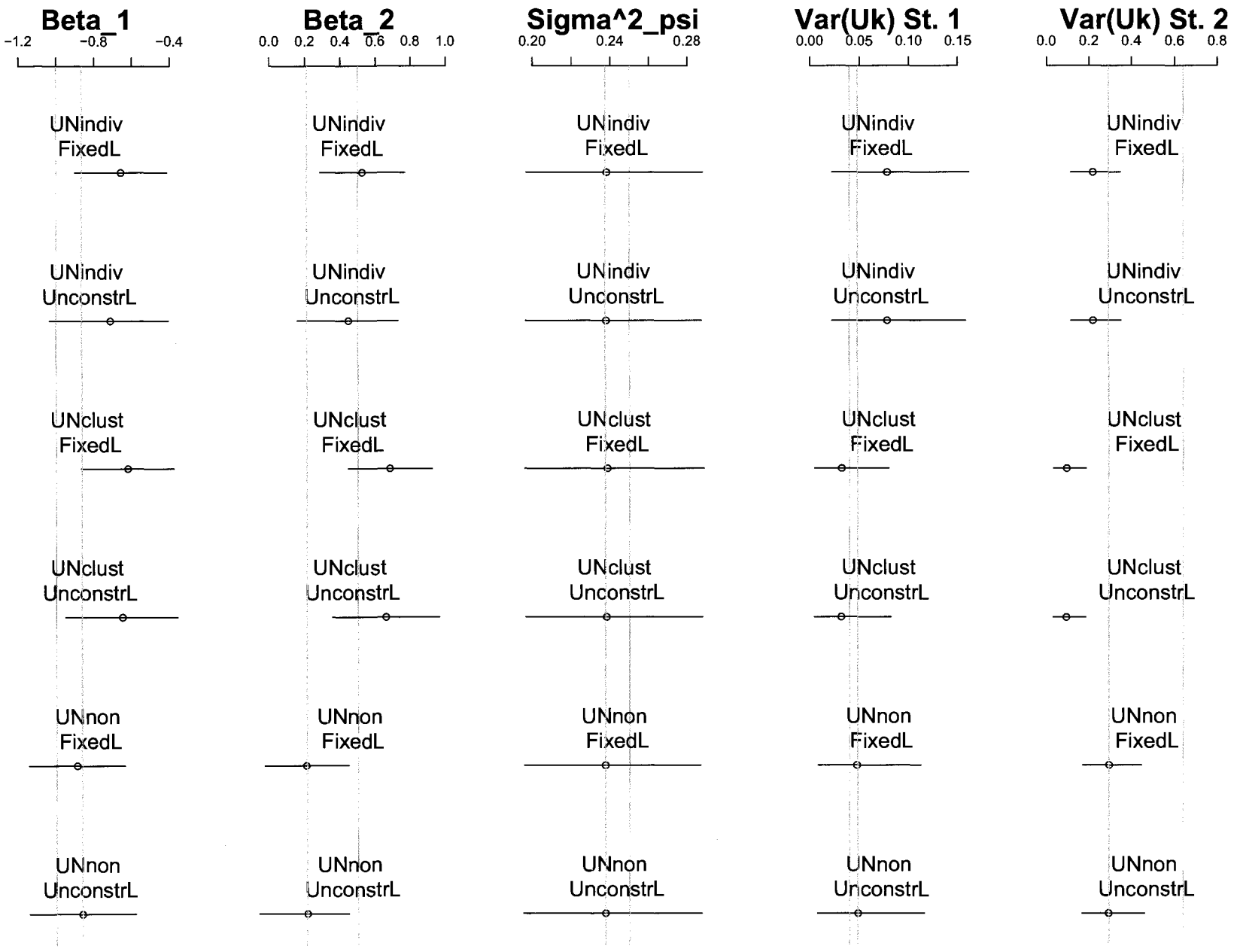

Figure 4.1: Unweighted GoM Results of Sampling Design Parameters (GoM Scores)

4.5.2 Weighted Results of Sampling Design Parameters (GoM Scores)

The weighted estimates of the sampling design parameters (GoM Scores) are in Figure 4.2. This figure is included for completeness, though it is difficult to interpret. There are two confounding effects in this comparison. The first difference is the difference between the estimates of the parameters of the polytomous logistic mixed-effects regression prior when λ is fixed versus when λ is unconstrained with an informative prior. The second difference is that the wEP weighting is different when λ is fixed and when λ is unconstrained. To see this consider the unweighted joint distribution from Equation 4.8. As discussed in the *Weighting based on the Estimated Parameter* subsection of Section 4.2.2, the $p(y_{kij}|m_{kijc}, \lambda_{cj})$ entry of the likelihood gets weights because the estimated parameter λ_{cj} is a group parameter. However, when λ_{cj} is fixed, then the term $p(y_{kij}|m_{kicj})$ no longer contains an estimated group parameter and does not get weights.

Given the confounding described above, there are a few items of note in Figure 4.2. First consider the fixed effects, or β parameter estimates. The parameter estimates when λ is unconstrained have a larger posterior spread then λ is fixed. This is reasonable since unconstrained λ's will contribute variability and the unconstrained λ wEP estimates contain weights in more places in the analysis (it is well known that weighted estimates have larger variances). The estimates of β_1 when λ is unconstrained have a lower mean than the estimates when λ is fixed. The means β_2 when λ is fixed and the estimates when when λ is unconstrained with an informative prior are much closer to each other than for β_1. The basic trend of the estimates when λ is fixed is as expected, with the informative cluster and individual sampling scheme estimates larger than the noninformative sampling

scheme. The same is true with the estimates when λ is unconstrained. I would expect the weighted estimates to have less bias in the wEP estimates than the unweighted estimates in Figure 4.1. This comparison is done in Figures 4.3 and 4.4.

Next consider the variance components. The estimates of σ^2_ψ when λ is unconstrained have similar posterior variance than the estimates when λ is fixed. The means of the estimates when λ is fixed are very close to the means of the estimates when λ is unconstrained. The general trend on the σ^2_ψ estimates is not expected. The estimates under informative cluster sampling are larger than the estimates under non-informative sampling and the estimates under informative individual sampling are larger than under informative cluster sampling. Consider the estimates of $\text{Var}(U_k)$ Stratum 1 and Stratum 2. The estimates when λ is unconstrained have larger posterior variance than the estimates when λ is fixed, as expected. The mean of the estimates of $\text{Var}(U_k)$ Stratum 1 when λ is fixed are close to the means when λ is unconstrained when the informative sampling does not affect this parameter (for the informative individual and non-informatve sampling schemes). When the informative sampling does affect this parameter (informative cluster sampling), the estimates when λ is unconstrained are closer to the true value. The same holds true for $\text{Var}(U_k)$ Stratum 2, except there is some difference between the estimates when λ is unconstrained and when λ is fixed in the non-informative sampling scheme, with the estimates when λ is fixed being closer to the true value.

Figure 4.2: Weighted Results of Sampling Design Parameters (GoM Scores)

4.5.3 Results of Sampling Design Parameters (GoM Scores) when λ is Fixed

Figure 4.3 compares the weighted and wEP weighted estimates of the sampling design paramters (GoM scores) when λ is fixed.

First consider the fixed-effects or β estimates. The general trend of the unweighted estimates is as expected, as noted in the description of Figure 4.1. The weighted estimate under the non-informative sampling scheme match the unweighted estimate, as expected. For the β_1 estimates under the informative cluster sampling scheme, the weighted estimate has more bias in the same direction as the unweighted estimate, which is not expected as the informative sampling of clusters increases the β estimates and the weighting should reduce the affect of the informative sampling. The weights do not appear to help under informative individual sampling, though they do not increase bias either. For β_2, the weighted estimates adjust the mean of the estimate in the correct direction, with some overcompensation.

Next, consider the variance components. The general trends in the unweighted estimates are in the description of Figure 4.1. The weighted estimate under noninformative sampling is biased low, as expected from the simulations from Chapter 3, for example the simulations summarized in Figure 3.3 when the estimated model is in Equation 3.7. The estimated model in Equation 3.7 contained sampling at random and the scaled 2 estimate of σ^2_ϵ showed negative bias. The weights used for the GoM simulations are similar to the scaled 2 weights. This underestimation of the random error variance is seen in all sampling schemes in Figure 4.3. The unweighted estimates of σ^2_ψ are similar across all sampling schemes (as discussed with Figure 4.1). The weighted estimates produce largest bias in the non-informative sampling scheme, and least bias in the informative individual sampling scheme. The weighted estimates of $\text{Var}(U_k)$ Stratum 1 are larger than the unweighted estimates. This is due to the small intra-class correlation (icc=$\frac{0.04}{0.25} = 0.1$) as described in Section

2.4.2. This overestimation of the random intercept was also seen in Chapter 2, for example in Figure 3.3 when the estimated model is in Equation 3.9. The model in Equation 3.9 contained informative individual sampling which should not have affected the estimate of the random intercept, however the intra-class correlation was reduced and additional bias was seen in the scaled 2 estimate of σ^2_{0k}. The weighted estimates of $\text{Var}(U_k)$ Stratum 2 are also larger than the corresponding unweighted estimates. The reason for this is not known. In Chapter 2, Simulation Set 12 in Figure 3.9 there was also positive bias of the scaled 2 weights on the estimate of the random intercept, σ^2_{0k}, when the intraclass correlation was not very small (icc$\approx \frac{5}{20} = 0.25$).

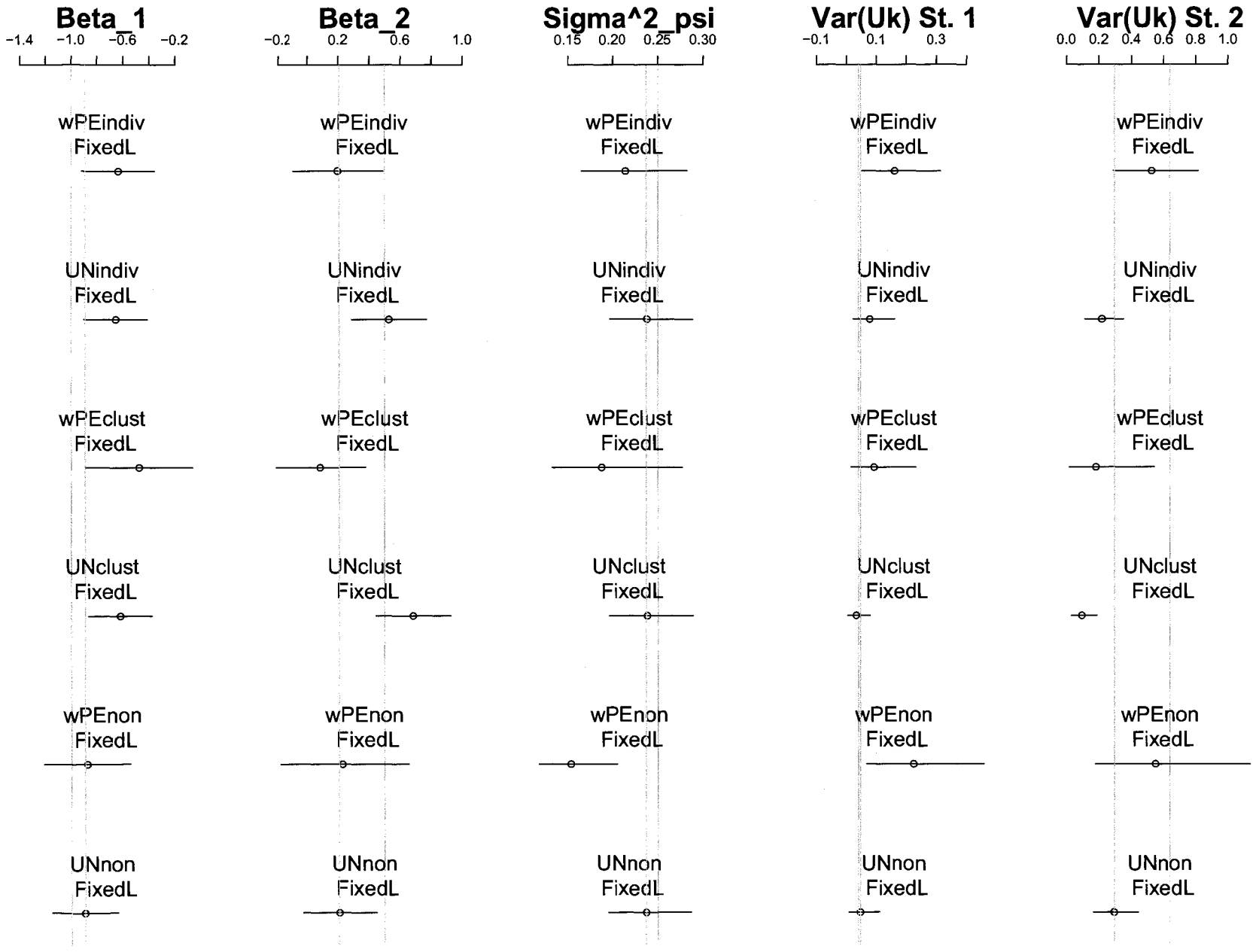

Figure 4.3: Weighted versus Unweighted GoM Results of Sampling Design Parameters (GoM Scores) when λ is Fixed

4.5.4 Results of Sampling Design Parameters (GoM Scores) when λ has an Informative Prior

Figure 4.4 compares the unweighted and wEP weighted estimates of the sampling design parameters (GoM scores) when λ is unconstrained with an informative prior.

First consider the fixed-effects or β estimates. The general trend of the unweighted estimates is as expected, as noted in the description of Figure 4.1. When there is informative cluster or individual sampling, the weighted estimates correctly compensate in the correct direction. For the estimate of β_1 in the non-informatve sampling scheme, the unweighted estimates produce negative bias. However, for the estimates of β_2 under non-informative sampling, the weighted and unweighted estimates have similar means. As expected the weighted estimates have larger posterior spread than the unweighted estimates.

Next consider the variance components. The behavior of the weighted and unweighted estimates of σ^2_ψ is similar to the behavior when λ is fixed as seen in Figure 4.3. The behavior of the estimates of $\mathrm{Var}(U_k)$ in Stratum 1 and Stratum 2 is also the same as when λ is fixed as seen in Figure 4.3, however the posterior spreads are larger than in the fixed λ case.

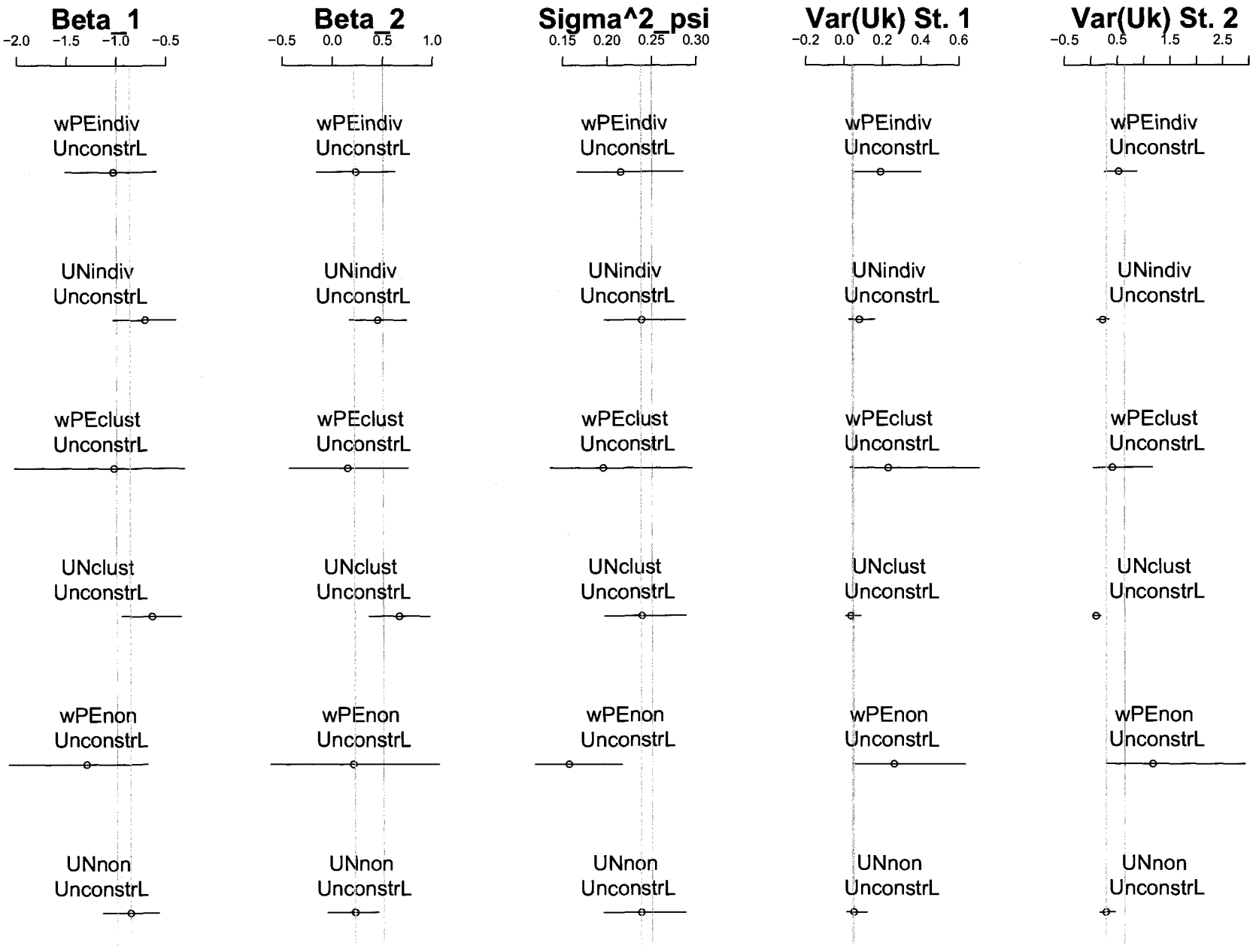

Figure 4.4: Weighted versus Unweighted GoM Results of Sampling Design Parameters (GoM Scores) with an Informative Prior on λ

4.5.5 Comparison of Weighted versus Unweighted Estimates of λ

Figure 4.5 compares the unweighted and wEP weighted estimates of λ. The scale on all the graphs is 0 to 1 as the λ parameters are probabilities.

The main feature of Figure 4.5 is the consistency of the means regardless of sampling scheme or type of weighting (unweighted or wEP weighted). I next highlight the estimates whose 0.025 and 0.975 quantiles either do not include the true value, or barely include it. These estimates include the unweighted and weighted estimates of $\lambda_{1,4}$ under the informative individual sampling scheme, the weighted estimate of $\lambda_{2,3}$ under the informative individual sampling scheme, and the weighted estimate of $\lambda_{2,4}$ under the informative cluster sampling scheme.

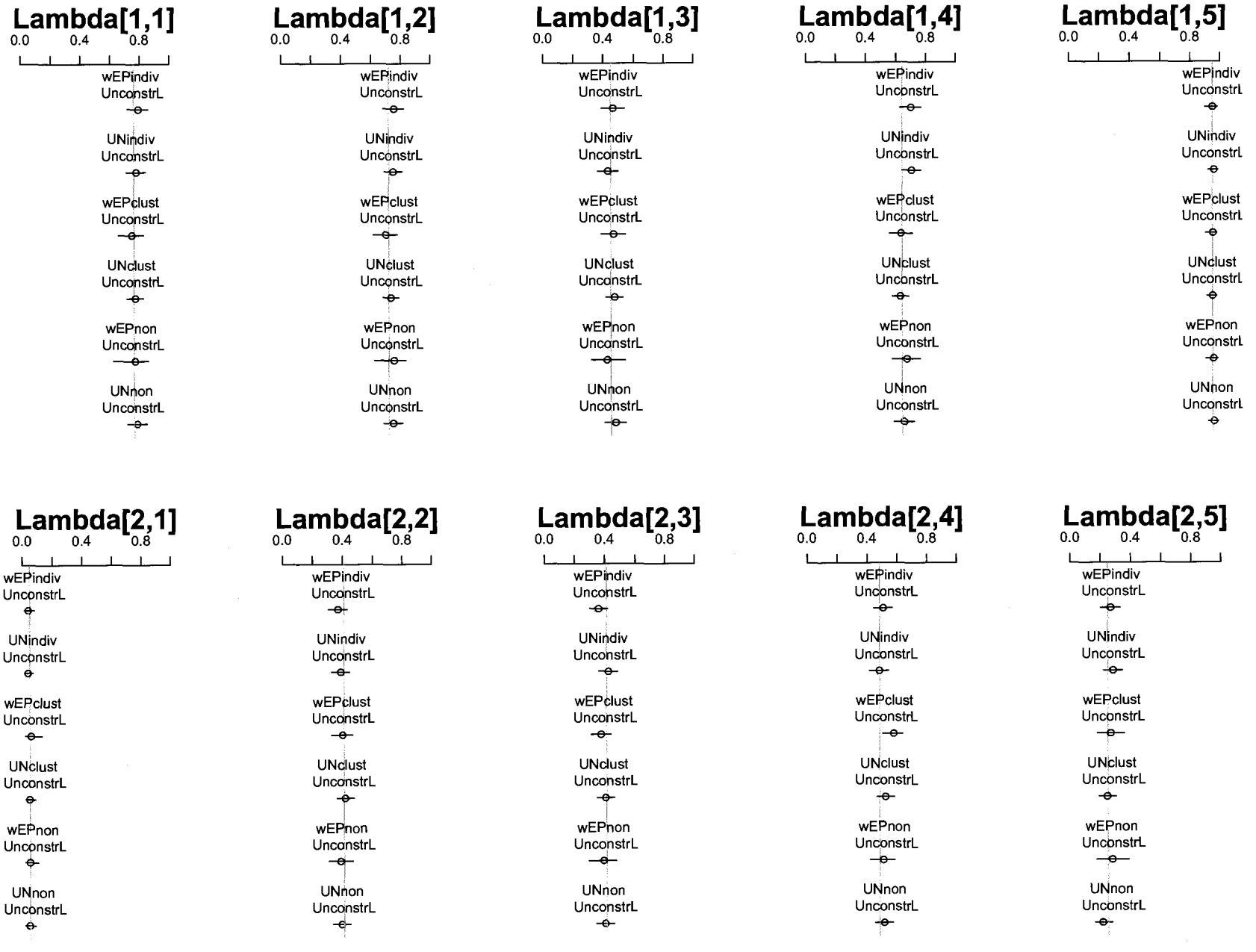

Figure 4.5: Weighted versus Unweighted Estimates of λ

4.6 Summary

The goals of this chapter are to 1) modify the GoM model to incorporate the sampling design, 2) insert weights into the modified GoM model and 3) analyze the performance of the new unweighted and weighted GoM model through a simulation study. A number of new contributions were made in supporting these three goals.

The original Dirichlet prior GoM model was modified to use a polytomous logistic mixed-effects regression prior. This prior allows incorporation of the dependencies in the GoM scores induced by the sampling design. Another advantage to this prior, as discussed in the future work, is that it can also easily analyze dependencies of longitudinal data.

The insertion of sampling weights expanded upon the PML method from Chapter 2. In addition, the new method, weighted based on the estimated parameter, introduced a principled type of weighting for complex analyses.

Lastly, the simulation study characterizes the performance of the new polytomous logistic mixed-effects regression prior and the weighting based on the estimated parameter. The simulations indicate that the effect of the sampling design and the effect of adding weights to the analysis strongly parallel the results from the LME simulation study in Chapter 3.

4.7 Appendices

4.7.1 PML Weighting of the GoM Model

The weighting method of PSHGR in Section 2.2.5 inserts the weights in the process of solving for the estimators. Using the Bayesian modeling and estimation techniques that are different from those used by PSHGR, it is difficult to update the PSHGR method for use on the GoM model. The weighting method of RHS in Section 2.2.3 creates a weighted likelihood, which is easily incorporated into the Bayesian GoM model structure.

A slight reparameterization of the prior on the U's is needed to incorporate the RHS weighting. Recall the variance structure of Ω from Equation 2.3 in Chapter 2. The model from Chapter 2 (and from GoM description above) assumes that the elements of U are ordered according to random effect, as in Equation 2.1. To incorporate the weighting of RHS, change the ordering of the Z and U matrices to be according to cluster instead of element, as in Equation 2.54. Allowing U_k to be the random effects corresponding to cluster k, the prior on U_{kc} for the GoM model becomes

$$\begin{aligned} U_{kc} &\sim \text{Normal}(0, \mathcal{O}), \quad , c = 1, \cdots, C-1 \\ & \qquad U_{k_1 c} \perp U_{k_2 c} \end{aligned} \tag{4.9}$$

where Equation 4.9 is a prior on each cluster for class c, $c = 1, ..., C-1$ and $\mathcal{O}$ is defined in Equation 2.3.

Consider the census joint distribution defined in Equation 4.7. Estimate the census joint dis-

tribution with the sample weighted joint distribution. Let

$$
\begin{aligned}
p_w(y,m,\psi,\lambda,\beta,U,\sigma^2_\psi,X,Z) \quad \propto \quad & \exp\left\{-\frac{1}{2}\sum_{c=1}^{C}(\beta_c-\mu_\beta)^T\Sigma_\beta^{-1}(\beta_c-\mu_\beta)\right\} \\
\times \quad & (\sigma^2_\psi)^{-(\frac{\nu}{2}+1)}\exp\left\{-\frac{\nu s^2_\psi}{2\sigma^2_\psi}\right\}\left[\prod_{c=1}^{C}\prod_{j=1}^{J}\lambda_{cj}^{\eta_{1cj}-1}(1-\lambda_{cj})^{\eta_{2cj}-1}\right] \\
\times \quad & \prod_{k=1}^{K_s}\prod_{i=1}^{n_k}\prod_{j=1}^{J}\prod_{c=1}^{C}\left[\frac{\exp(\psi_{kic})}{\sum_{c_1=1}^{C}\exp(\psi_{kic_1})}\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{1-y_{kij}}\right]^{m_{kijc}w_{ki}} \\
\times \quad & \prod_{k=1}^{K_s}\left[\prod_{i=1}^{n_k}\left(\prod_{c=1}^{C-1}(\sigma^2_\psi)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2_\psi}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\}\right)^{w_{i|k}}\right. \\
\times \quad & \left.\exp\left\{-\frac{1}{2}U_{kc}^T\mathcal{O}^{-1}U_{kc}\right\}\right]^{w_k} \qquad (4.10)
\end{aligned}
$$

where K_s is the number of sampled clusters and n_k is the number of sampled individuals in cluster k. The weights are not inserted on the prior distributions of β, σ^2_ψ or λ and are inserted in the likelihood conditional distributions of $y_{ki}|_{m_{kijc},\lambda_{cj}}$ and $m_{kijc}|\psi_{kic}$. The weighting of the ψ's mimics the RHS weighting, which weights both the $\psi_{kic}|\beta,U,\sigma^2_{psi}$ and U distributions. The incorporation of sampling weights on the prior of U is reasonable if the prior on U only contains priors on the random effects for the sampled clusters. This weighting of the sample joint density propagates to

the complete conditionals corresponding to the polytomous regression parameters as follows:

$$
\begin{aligned}
p_w(\beta_c|-) &\propto \exp\left\{-\frac{1}{2}(\beta_c-\mu_c)^T\Sigma_\beta^{-1}(\beta_c-\mu_c)\right\}\prod_{k=1}^{K_s}\prod_{i=1}^{n_k}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\}^{w_{ki}} \\
&\sim \text{Normal}(\mu_1,\Sigma_1) \\
&\quad \mu_1=\left(\Sigma_\beta^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{Ks}\sum_{i=1}^{n_k}w_{ki}X_{ki}^TX_{ki}\right)^{-1}\left(\Sigma_\beta^{-1}\mu_c+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}X_{ki}^T(\psi_{kic}-Z_{ki}U_c)\right) \\
&\quad \Sigma_1=\left(\Sigma_\beta^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}X_{ki}^TX_{ki}\right)^{-1} \\
p_w(U_c|-) &\propto \prod_{k=1}^{K}\left[\prod_{i=1}^{n_k}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_{kc})^2\right\}^{w_{i|k}}\exp\left\{-\frac{1}{2}U_{kc}^T\mathcal{O}^{-1}U_{kc}\right\}\right]^{w_k} \\
&\sim \text{Normal}(\mu_2,\Sigma_2) \\
&\quad \mu_2=\left(\sum_{k=1}^{K}w_k\mathcal{O}^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}Z_{ki}^TZ_{ki}\right)^{-1}\left(\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}Z_{ki}^T(\psi_{kic}-X_{ki}\beta_c)\right) \\
&\quad \Sigma_2=\left(\sum_{k=1}^{K}w_k\mathcal{O}^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}Z_{ki}^TZ_{ki}\right)^{-1} \\
p_w(\sigma_\psi^2|-) &\propto (\sigma_\psi^2)^{-(\frac{\nu}{2}+1)}\exp\left\{-\frac{\nu s_\psi^2}{2\sigma_\psi^2}\right\}\prod_{k=1}^{K_w}\prod_{i=1}^{n_k}\prod_{c=1}^{C-1}(\sigma_\psi^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kc}-X_k\beta_c-Z_kU_c)^2\right\}^{w_{ki}} \\
&\sim \text{Scaled Inv }\chi^2\left(\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}(C-1)+\nu,\frac{\left(\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}\sum_{c=1}^{C-1}w_{ki}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right)+\nu s_\psi^2}{\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}(C-1)+\nu}\right)
\end{aligned}
$$

The complete conditionals for the augmented data and the pure response probabilities are,

$$
\begin{aligned}
p_w(\lambda_{cj}|-) &\propto \prod_{k=1}^{K_s}\prod_{i=1}^{n_k}\left[\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right]^{m_{kijc}w_{ki}}\lambda_{cj}^{\eta_{1cj}-1}(1-\lambda_{cj})^{\eta_{2cj}-1} \\
&\sim \text{Beta}\left(\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}y_{kij}m_{kijc}w_{ki}+\eta_{1cj},\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}(1-y_{kij})m_{kijc}w_{ki}+\eta_{2cj}\right) \\
p_w(m_{kij}|-) &\propto \prod_{c=1}^{C}\left[\frac{\exp\{\psi_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right]^{m_{kijc}w_{ki}} \\
&\sim \text{Multinomial}\left(1,\left[\frac{\exp\{\psi_{ki1}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{1j}^{y_{kij}}(1-\lambda_{1j})^{1-y_{kij}}\right]^{w_{ki}},\cdots,\right. \\
&\quad \left.\left[\frac{\exp\{\psi_{kiC}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{Cj}^{y_{kij}}(1-\lambda_{Cj})^{1-y_{kij}}\right]^{w_{ki}}\right)
\end{aligned}
$$

Finally, a Metropolis step is needed for ψ,

$$\begin{aligned} p_w(\psi_{kic}|-) &\propto \prod_{c=1}^{C}\prod_{j=1}^{J}\left[\frac{\exp\{\psi_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\right]^{m_{kijc}w_{ki}} \\ &\times \exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\} \end{aligned}$$

Let the Jumping distribution be Normally distributed, with the mean at the previous MCMC value and a variance of $\sigma^2_{\psi\text{jmp}}$. The acceptance ratio is

$$\begin{aligned} r_{(w)\psi_{kic}} &= \frac{p_w(\psi^*_{kic}|-)}{p_w(\psi^{(r)}_{kic}|-)} \\ &= \prod_{c=1}^{C}\prod_{j=1}^{J}\left[\frac{\exp\{\psi^*_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi^*_{kic_1}\}}\frac{\sum_{c_1=1}^{C}\exp\{\psi^{(r)}_{kic_1}\}}{\exp\{\psi^{(r)}_{kic}\}}\right]^{m_{kijc}w_{ki}} \\ &\times \exp\left\{-\frac{1}{2\sigma_\psi^2}\left[(\psi^*_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2-(\psi^{(r)}_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right]\right\} \\ &= \prod_{c=1}^{C}\left[\frac{g^*_{kic}}{g^{(r)}_{kic}}\right]^{\sum_{j=1}^{J}m_{kijc}w_{ki}} \exp\left\{-\frac{1}{2\sigma_\psi^2}\left[(\psi^*_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2-(\psi^{(r)}_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right]\right\} \end{aligned}$$

where $g_{kic} = \frac{\exp\{\psi_{ic}\}}{\sum_{c_1=1}^{C}\psi_{kic_1}}$ where $\psi_{kiC} = 0$. These complete conditionals and Metropolis-Hasting steps are implemented using MCMC algorithms.

4.7.2 Complete Conditional Weighting

To add weights to the complete conditionals, take the census complete conditionals from the unweighted GoM derivation, and add sampling weights to estimate them with sample complete conditionals. The subscript wCC below denotes the result from weighting the complete condtionals. The weighted complete conditionals for the polytomous regression parameters are the same as above,

$$
\begin{aligned}
p_{wCC}(\beta_c|-) &\propto \exp\left\{-\frac{1}{2}(\beta_c-\mu_c)^T\Sigma_\beta^{-1}(\beta_c-\mu_c)\right\}\prod_{k=1}^{K_s}\prod_{i=1}^{n_k}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\}^{w_{ki}} \\
&\sim \text{Normal}(\mu_1,\Sigma_1) \\
&\quad \mu_1=\left(\Sigma_\beta^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{Ks}\sum_{i=1}^{n_k}w_{ki}X_{ki}^TX_{ki}\right)^{-1}\left(\Sigma_\beta^{-1}\mu_c+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}X_{ki}^T(\psi_{kic}-Z_{ki}U_c)\right) \\
&\quad \Sigma_1=\left(\Sigma_\beta^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}X_{ki}^TX_{ki}\right)^{-1} \\
p_{wCC}(U_c|-) &\propto \prod_{k=1}^{K}\left[\prod_{i=1}^{n_k}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_{kc})^2\right\}^{w_{i|k}}\exp\left\{-\frac{1}{2}U_{kc}^T\mathcal{O}^{-1}U_{kc}\right\}\right]^{w_k} \\
&\sim \text{Normal}(\mu_2,\Sigma_2) \\
&\quad \mu_2=\left(\sum_{k=1}^{K}w_k\mathcal{O}^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}Z_{ki}^TZ_{ki}\right)^{-1}\left(\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}Z_{ki}^T(\psi_{kic}-X_{ki}\beta_c)\right) \\
&\quad \Sigma_2=\left(\sum_{k=1}^{K}w_k\mathcal{O}^{-1}+\frac{1}{\sigma_\psi^2}\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}Z_{ki}^TZ_{ki}\right)^{-1} \\
p_{wCC}(\sigma_\psi^2|-) &\propto (\sigma_\psi^2)^{-(\frac{\nu}{2}+1)}\exp\left\{-\frac{\nu s_\psi^2}{2\sigma_\psi^2}\right\}\prod_{k=1}^{K_w}\prod_{i=1}^{n_k}\prod_{c=1}^{C-1}(\sigma_\psi^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kc}-X_k\beta_c-Z_kU_c)^2\right\}^{w_{ki}} \\
&\sim \text{Scaled Inv }\chi^2\left(\nu_1,s_1^2\right) \\
&\quad \nu_1=\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}(C-1)+\nu \\
&\quad s_1^2=\frac{\left(\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}\sum_{c=1}^{C-1}w_{ki}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right)+\nu s_\psi^2}{\sum_{k=1}^{K_s}\sum_{i=1}^{n_k}w_{ki}(C-1)+\nu}
\end{aligned}
$$

The complete conditional for λ_{cj} remains the same, and the complete conditional for m_{kijc} changes as described above,

$$\begin{aligned}
p_{wCC}(\lambda_{cj}|-) &\propto \prod_{k=1}^{K_s}\prod_{i=1}^{n_k}\left[\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right]^{m_{kijc}w_{ki}} \lambda_{cj}^{\eta_{1cj}-1}(1-\lambda_{cj})^{\eta_{2cj}-1} \\
&\sim \text{Beta}\left(\sum_{k=1}^{K_s}\sum_{i=1}^{n_k} y_{kij}m_{kijc}w_{ki}+\eta_{1cj}, \sum_{k=1}^{K_s}\sum_{i=1}^{n_k}(1-y_{kij})m_{kijc}w_{ki}+\eta_{2cj}\right) \\
p_{wCC}(m_{kij}|-) &\propto \prod_{c=1}^{C}\left[\frac{\exp\{\psi_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{(1-y_{kij})}\right]^{m_{kijc}} \\
&\sim \text{Multinomial}\left(1, \frac{\exp\{\psi_{ki1}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{1j}^{y_{kij}}(1-\lambda_{1j})^{1-y_{kij}}, \cdots, \right. \\
&\quad \left. \frac{\exp\{\psi_{kiC}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\lambda_{Cj}^{y_{kij}}(1-\lambda_{Cj})^{1-y_{kij}}\right)
\end{aligned}$$

The Metropolis-Hastings step for ψ_{kic} is the same as the unweighted case,

$$\begin{aligned}
p_{wCC}(\psi_{kic}|-) &\propto \prod_{c=1}^{C}\prod_{j=1}^{J}\left[\frac{\exp\{\psi_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi_{kic_1}\}}\right]^{m_{kijc}} \\
&\times \exp\left\{-\frac{1}{2\sigma_\psi^2}(\psi_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right\}
\end{aligned}$$

Let the Jumping distribution be Normally distributed, with the mean at the previous MCMC value and a variance of $\sigma^2_{\psi\text{jmp}}$. The acceptance ratio is

$$\begin{aligned}
r_{(wCC)\psi_{kic}} &= \frac{p(\psi^*_{kic}|-)}{p(\psi^{(r)}_{kic}|-)} \\
&= \prod_{c=1}^{C}\prod_{j=1}^{J}\left[\frac{\exp\{\psi^*_{kic}\}}{\sum_{c_1=1}^{C}\exp\{\psi^*_{kic_1}\}}\frac{\sum_{c_1=1}^{C}\exp\{\psi^{(r)}_{kic_1}\}}{\exp\{\psi^{(r)}_{kic}\}}\right]^{m_{kijc}} \\
&\times \exp\left\{-\frac{1}{2\sigma_\psi^2}\left[(\psi^*_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2-(\psi^{(r)}_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right]\right\} \\
&= \prod_{c=1}^{C}\left[\frac{g^*_{kic}}{g^{(r)}_{kic}}\right]^{\sum_{j=1}^{J}m_{kijc}} \exp\left\{-\frac{1}{2\sigma_\psi^2}\left[(\psi^*_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2-(\psi^{(r)}_{kic}-X_{ki}\beta_c-Z_{ki}U_c)^2\right]\right\}
\end{aligned}$$

By construction, these complete conditionals and Metropolis-Hastings steps insert weights only when using sample quantities to estimate finite population quantities.

One problem with this wCC weighting that some components to the posterior (or log likelihood times the prior) are treated differently in different complete conditionals. For example, in $p_{wCC}(\lambda_{cj}|-)$, the $\left[\lambda_{cj}^{y_{kij}}(1-\lambda_{cj})^{1-y_{kij}}\right]$ term from the posterior is weighted. However, the same term in $p_{wCC}(m_{kij}|-)$ is not weighted. Treating a component from the posterior differently in different complete conditionals appears unprincipled.

4.7.3 Computer Code

The GoM code used for the results of this chapter started with the code from Elena Erosheva's thesis, see Erosheva (2002), and was modified by Cyrille Joutard. I then modified that code to include the polytomous logistic random-effects prior and the wEP weighting scheme. This code may be found at `http://stat.cmu.edu` under the `Recent PhD Theses` link. The c-code uses the VMR library, downloaded from `http://www.stat.cmu.edu/~hseltman/`. It is in the Computer Programming, C/C++ section. The code uses the `IMSL` library, available from Visual Numerics at `http://www.vni.com` for a fee. The compilation instructions are commented in the beginning of the code. Along with the ode are sample input files and the corresponding output file.

Chapter 5

Conclusions and Future Research

5.1 Contributions

The way in which the sampling design should be incorporated into model-based analyses has triggered much controversy between design- and model-based analysts. In this thesis, I investigated when the weights benefit the analysis by reducing bias and when they do not benefit the analysis. I investigated model-based ways of incorporation of the sampling design by utilizing linear mixed-effect models and their flexible variance structure. I also investigated principled ways of inserting sampling weights into complex model-based analyses. In the process of these investigations, I made contributions to the existing literature.

LME Contributions (Methodological and Analytic). The methodological contributions started by describing three competing approaches to inserting weights into linear mixed-effects models, Rabe-Hesketh and Skrondal (2006), Korn and Graubard (2003) and Pfeffermann et al. (1998), showing their common underlying framework and where each method makes unique decisions. While pseudo-maximum likelihood (PML) sounds good in principle, the three methods all utilize PML, inserting weights in different parts of the analysis and developing competing estima-

tors. This investigation enforced the ad-hoc nature of many PML analyses. I compared the three different approaches towards variance estimation, the sandwich estimator of the variance, design-based approximation of the variance and survey adjusted jackknife estimation of the variance.

The analytic contributions are derived from an extensive simulation study involving 12 simulation sets. There are five main conclusions from the simulations. 1) Differences in the PSHGR and RHS methods are very small and due to numerical instabilities in the estimation procedures. 2) The sandwich estimator provides better coverage, specifically in the presence of model misspecification. 3) When there is model misspecification that does not induce informative sampling, weighted estimates do not reduce the bias of the estimators. 4) When there is informative sampling, the weighted estimators do reduce the bias of the point estimates, though they do not eliminate it. 5) The different scalings of the weights have different characteristics. The unweighted estimate has the smallest variance, though are biased in the presence of informative sampling. The weighted unscaled estimates correct for the bias in the fixed effects, but produce more bias in the random effects. The scaled 1 weightings correct for the bias in the fixed effects and overcorrect for the bias in the random effects. The scaled 2 estimates are in-between the unscaled and scaled 2 estimates.

GoM Contributions (Methodological and Analytic). The methodological contribution for the GoM model involves the incorporation of the polytomous logistic mixed-effects regression prior. This prior models the complex dependencies induced by the sampling design and can easily be modified to reflect dependencies of longitudinal data. I also developed a principled method of weighting, called weighting based on the estimated parameter. This allows sampling weights to be incorporated into model-based analyses and without the ad-hoc nature that appears in PML.

The analytic contributions are derived from a simulation study evaluating the polytomous logistic mixed-effects regression prior and comparing the unweighted and weighted based on the estimated parameter analyses. These simulations demonstrate strong parallels with the simulation studies in Chapter 3.

5.2 Future Work

LME and GoM Models. An interesting connection between the LME simulations and the GoM simulations would investigate a Bayesian formulation of the LME model analyzed via MCMC. Comparisons of the unweighted versus weighted analyses in this context can provide more insight to the behavior of the weights, especially the weighting based on the estimated parameter.

The MCMC algorithms for the estimation of the GoM parameters need to be improved. The simulations in Chapter 4 include an informative prior on σ^2_ψ, which was needed to control drifting of this parameter. The reason for the drifting of the parameter estimate needs to be better understood, especially if the reason is due to non-identifiability issues. The β parameters are also very sensitive to drifting and this is much worse when the prior on σ^2_ψ is noninformative. Finally, the sensitivity of rotational indeterminancy to the strength of the informativeness of the prior on λ needs to be better understood. Sensitivity analyses that vary the informative prior can provide insights into the robustness of the model.

Once the MCMC algorithms are improved, the National Long Term Care Survey data should be analyze using the GoM model. A detailed analysis comparing the different numbers of latent classes, similar to the analyses in Erosheva (2002), should be done to see how the weights affect the distributions of the GoM scores. The addition of the polytomous logistic mixed-effects regression prior allows for the longitudinal nature of the NLTCS data by incorporating each person within a cluster in the structure of the Z matrix.

Survey Sampling and Sampling Weights. The next step for the sampling weights is to investigate a weighted longitudinal analysis within the NLTCS data. Large national surveys such as NLTCS produce may sets of weights for cross section and longitudinal analyses. Investigation on how to incorporate the different types of weights in a longitudinal analysis is needed.

Bibliography

Airoldi, E. M., Fienberg, S. E., Joutard, and C. Love, T. M. (2005). Hierarchical Bayesian mixed-membership models and latent pattern discovery. *International Journal of Parallel, Emergent and Distributed Systems*, 00(00):1–12.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics – Theory and Methods*, 35:439 – 460.

Binder, D. and Roberts, G. (2006). Approaches for analyzing survey data: a discussion. In *Proceedings of the Survey Research Methods Section, American Statistical Association (2006)*, pages 2771–2778.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51(3):279 – 292.

Bradley, R. and Gart, J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika*, 49(1/2):205–214.

Chambless, L. and Boyle, K. (1985). Maximum likelihood methods for complex survey data: Logistic regression and discrete proportional hazards models. *Communication in Statistics, Theory and Methods*, 14(6):1377–1392.

Cowles, M. and Carlin, B. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 92(434):883–904.

Down, K., Kinsey, S., Wheeless, S., Thissen, R., Richardson, J., Mierzwa, F., and Biemer, P. (2002). *National Survey of Child and Adolescent Well-Being (NSCAW): Introduction to the Wave 1 General and Restricted Use Releases.* National Data Archive on Child Abuse and Neglect, Cornell University Ithaca, NY 14853.

Erosheva, E. (2002). *Grade of Membership and Latent Structure Models with Application to Disability Survey Data.* PhD thesis, Carnegie Mellon University, Department of Statistics.

Erosheva, E. (2005). Comaring latent structures of the grade of memberhisp, rasch and latent class models. *Psychometrika*, 70(4):619–628.

Fienberg, S. E. (1980). The measurement of crime victimization: Prospects for panel analysis of a panel survey. *Statistician*, 29:313–350.

Fienberg, S. E. (1989). Modeling considerations: Discussion from a modeling perspective. In Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M. P., editors, *Panel Surveys*, pages 512–539. Wiley.

Fisher, R. A. (1935). *Design of Experiments.* Oliver & Boyd.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164.

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

Godambe, V. P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society B*, 17:269–278.

Godambe, V. P. (1966). A new approach to sampling from finite populations. *Journal of the Royal Statistical Society B*, 28:310–328.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54:127–138.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73(1):43 – 56.

Graubard, B. I. and Korn, E. L. (1995). Modeling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, 5:263–281.

Hansen, M. Madow, W. and Tepping, B. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78:776–793.

Hartley, H. O. and Sielken, R. L. (1975). A "super-population viewpoint" for finite population sampling. *Biometrics*, 31(2):411–422.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31:1109–1144.

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*, pages 512–539. Wiley.

Holt, D. and Smith, T. (1979). Post-stratification. *Journal of the Royal Statistical Society A*, 142:33–46.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Huang, R. and Hidiroglou, M. (2003). Design consistent estimators for a mixed linear model on survey data. In *Proceedings of the Survey Research Methods Section, American Statistical Association (2003)*, pages 1897–1904.

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis.* Simon & Schuster.

Kalton, G. (1968). Invited discussion to smith, t. m. f., present position and potential developments: Some personal views: Sample surveys. *Journal of the Royal Statistical Society A*, 147:208–221.

Kalton, G. (1989). Modeling considerations: Discussion from a survey sampling perspective. In Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M., editors, *Panel Surveys*, pages 575–585. Wiley.

Kaufman, G. M. and Press, S. J. (1973). Bayesian factor analysis. Technical Report 7322, Center for Mathematical Studies in Business and Economics, Department of Economics and Graduate School of Business.

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

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.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). *SAS System for Mixed Models.* SAS Publishing.

Little, R. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78:596–604.

Little, R. (1991). Inference with survey weights. *Journal of Official Statistics*, 7:405–424.

Little, R. (1993). Post-stratification: A modeler's perspective. *Journal of the Americal Statistical Association*, 88:1001–1012.

Little, R. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the Americal Statistical Association*, 99:546–556.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data.* Wiley.

Lohr, S. and Liu, J. (1994). A comparison of weighted and unweighted analyses in the ncvs. *Journal of Quantitative Criminology*, 10:343–360.

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

Longford, N. (2004). Small area estimation with complex sampling design. In *Enhancing Small Area Estimation Techniques to meet European Needs: Project Reference Volume, Vol.2: Explanatory Appendices.*

Marini, M., Li, X., and Fan, P.-L. (1996). Characterizing latent structure: Factor analytic and grade of membership models. *Sociological Methodology*, 26:133–164.

Mislevy, R. J. and Sheehan, K. M. (1989a). Information matrices in latent-variable models. *Journal of Educational Statistics*, 14(4):335–350.

Mislevy, R. J. and Sheehan, K. M. (1989b). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, 54(4):661–679.

Molina, E. A., Smith, T. M. F., and Sugden, R. A. (2001). Modelling overdispersion for complex survey data. *International Statistical Review*, 69(3):373–384.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–625.

NLTCS (1988). *Overview and Use of the Public Use Data Files of the 1982 and 1984 National Long Term Care Surveys*. Duke University Center for Demographic Studies, 2117 Campus Drive, Durham NC 27706.

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.

Pawitan, Y. (2001). *In all Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61(2):317 – 337.

Pfeffermann, D., Moura, F. A. D. S., and Silva, P. L. D. N. (2006). Multilevel modelling under informative sampling. *Biometrika*, 93:943–959.

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60(1):23 – 40.

Potthoff, R., Woodbury, M., and Manton, K. (1992). "equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association*, 87(418):838–396.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A*, 169:805–827.

Raftery, A. E. and Lewis, S. (1992). How many iterations in the gibbs sampler? In Bernardo, J. M., Berger, J., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, pages 763–773. WOxford University Press.

Rowe, D. B. (2001). A model for Bayesian factor analysis with jointly distributed means and loadings. Technical Report 1108, Division of the Humanities and Social Sciences, California Institute of Technology, Social Science Working Paper.

Royall, R. (1968). An old approach to finite population sampling theory. *Journal of the American Statistical Association*, 63:1269–1279.

Royall, R. (1976). Likelihood functions in finite population sampling theory. *Biometrika*, 63:605–614.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. In Bernardo, J. M., Degroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics 2*, pages 463–472. Elsevier Science Publishers.

Sarndal (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, 5:27–52.

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

Searle, S., Casella, G., and McCulloch, C. (1992). *Variance Components.* John Wiley & Sons, Inc.

Searle, S. R. (1982). *Matrix Algebra Useful for Statistics.* John Wiley & Sons, Inc.

Skinner, C. J. (1989a). 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.

Skinner, C. J. (1989b). Introduction to part a. In Skinner, C. J., Holt, D., and Smith, T. M. F., editors, *Analysis of Complex Surveys*, pages 59–87. Wiley.

Smith, T. M. F. (1988). To weight or not to weight, that is the question. In Bernardo, J. M., Degroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics 3*, pages 437–451. Oxford University Press.

Smith, T. M. F. and Sugden, R. A. (1988). Sampling and assignment mechanisms in experiments, surveys and observational studies. *International Statistical Review*, 56(2):165–180.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis*. Sage Publications.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling*, 9(4):475–502.

Sugden, R. A. (1979). Inference on symmetric functions of exchangeable populations. *Journal of the Royal Statistical Society. Series B*, 41(2):269–273.

Sugden, R. A. (1985). A Bayesian view of ignorable designs in survey sampling inference. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics 2*, pages 751–754. Elsevier Science Publishers.

Sugden, R. A. and Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71(3):495–506.

Thomas, D. R. and Cyr, A. (2002). Applying item response theory methods to complex survey data. In *Proceedings of the Survey Methods Section*, pages 17–25. Statistical Society of Canada.

Thomas, N. (2000). Assessing model sensitivity of the imputations methods used in the national assessment of educational progress. *Journal of Educational and Behavioral Statistics*, 25(4):351–371.

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

Vinovskis, M. (1998). *Overseeing the Nation's Report Card: The Creation and Evolution of the National Assessment Governing Board (NAGB)*. School of Public Policy, University of Michigan.

von Davier, M., Sinharay, S., Oranje, A., and Beaton, A. (2007). The statistical procedures used in national assessment of educational progress: Recent developments and future directions. In *Handbook of Statistics:Psychometrics*.

Weisberg, S. (2005). *Applied Linear Regression*. Wiley.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag.

Woodbury, M. and Manton, K. (1989). Grade-of-membership analysis of depression-related psychiatric disorders. *Sociological Methods and Research*, 18:126–163.