Contribution to the Analysis of Complex Survey
Data and Cluster-Correlated Biological Data
Using Inverse Sampling

by

Emmanuel Benhin, B.Sc., M.Sc.

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of Doctor of Philosophy

Ottawa-Carleton Institute of Mathematics and Statistics

Carleton University, Ottawa, Ontario

June 2004

Copyright © 2004 by Emmanuel Benhin

Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-97811-7

Our file Notre référence

ISBN: 0-612-97811-7

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

This thesis addresses two main areas of statistical application. The first part of the thesis focuses on the analysis of complex survey data using inverse sampling. In the second part of the thesis some new ideas of analyzing cluster-correlated biological data are discussed.

Application of classical statistical methods to complex survey data without making allowance for the survey design features can lead to erroneous inferences. Many methods have been developed that account for the survey design, but these methods require additional information such as survey weights, design effects or cluster identification for microdata. Hinkins, Oh and Scheuren (1997) proposed inverse sampling as an alternative approach to analyzing complex survey data by undoing the complex survey data structures so that standard methods can be applied. Part I of this thesis discusses various inverse sampling schemes, some theory of inverse sampling and explore its strengths and weaknesses. We propose an estimating equations approach for handling complex parameters, such as ratios and “census” regression parameters, which naturally extends to poststratification estimation. We also explore the use of inverse sampling analyses of categorical complex survey data. Inverse sampling methods for testing hypotheses using Wald and Quasi-score tests for complex survey data are also studied.

The first part of Part II of this thesis discusses analyses of cluster-correlated biological data. Most classical methods for analyzing cluster-correlated biological data implicitly assume that the cluster sizes are ignorable. When this assumption

fails, these methods may be invalid. Within-cluster resampling proposed by Hoffman et al. (2001) provides a simple but computationally intensive method which is valid whether cluster sizes are ignorable or nonignorable. This method, however, can lead to unstable parameter estimators or variance estimators for small number of clusters. We propose two methods to rectify these shortcomings: combined estimating equation and mean estimating equation. We present some theory of the proposed methods and explore potential areas of application.

In the second part of Part II of this thesis, we discuss some methods of analyzing cluster-correlated binary response data. Conditional logistic regression (CLR) method for analyzing cluster-correlated binary response data implicitly assumes that the dependence arising from random cluster effects is a nuisance and all unmeasured cluster-specific risk factors are aggregated into a cluster-specific baseline. It is however, invalid when these assumptions fail. Rieger and Weinberg (2002) proposed within-cluster paired resampling method which remains valid when the assumptions hold or fail. The Rieger and Wienberg method can be computationally very intensive and can lead to unstable parameter estimators or variance estimators for small number of clusters. We propose an alternative method to rectify these shortcomings: mean conditional estimating equation. Some properties and basic theory of the proposed method are discussed.

The final chapter of Part II of this thesis discusses estimation methods for analyzing data having two types of correlation: within-cluster correlation and longitudinal correlation where the cluster sizes may be ignorable or nonignorable. We first study the effect on the efficiency of the regression parameter estimators when the assumed

working correlation structures depart from the true correlation structure. Secondly, we demonstrate that methods that implicitly assume ignorable cluster sizes can lead to asymptotically invalid inferences when this assumption fails. We study properties of proposed alternative approaches that lead to asymptotically valid inferences when cluster sizes are ignorable or nonignorable.

Acknowledgements

First, I would like to thank my supervisor Professor J.N.K. Rao for his advice, encouragement, direction and tremendous patience during the course of this work.

I would like to express my sincere appreciation and thanks to my Church family of St. George's Anglican Church, Ottawa for their prayers, encouragement and moral support.

Finally a very special thanks to my wife Agnes for her unwavering encouragement and support throughout this research work.

Contents

Abstract	ii
Acknowledgements	v
Contents	vi
List of Tables	xi
1 Introduction	1
1.1 Results of Chapter 2	3
1.1.1 Inverse Sampling Algorithms	3
1.1.2 Estimation of Simple Parameters	4
1.2 Results of Chapter 3	6
1.2.1 Estimation of Complex Parameters	6
1.2.2 Some Applications	6
1.3 Results of Chapter 4	7
1.4 Results of Chapter 5	8
1.5 Results of Chapter 6	9
1.6 Conclusions and Future Research	10

I	Analysis of Complex Survey Data Using Inverse Sampling	11
2	Basic Theory of Inverse Sampling	12
2.1	Introduction	12
2.2	Inverse Sampling Algorithms	14
2.2.1	Stratified Simple Random Sampling	16
2.2.2	One-Stage Cluster Sampling	17
2.2.3	Two-Stage Cluster Sampling	20
2.2.4	Stratified Two-Stage Sampling	21
2.3	Basic Properties of Inverse Sampling	22
2.4	Estimating Linear Population Parameters	26
2.4.1	Exact Matching	26
2.4.2	Estimation of a Population Total Using PPS	31
2.4.3	Approximate Matching	36
2.5	Improving the Efficiency of $\hat{Y}_g$	39
2.6	Inverse-Sampling Generalized Regression Estimator of a Total	41
2.6.1	Inverse-Sampling Difference Estimator	41
2.6.2	Inverse-Sampling Regression Estimator	43
2.7	Estimating Quadratic Population Parameters	46
3	Estimation of Complex Parameters	52
3.1	Estimating Equations Approach	52
3.2	Basic Properties of the CEE Method	53

3.2.1	Full-sample estimating equations	53
3.2.2	Separate estimating equations	55
3.2.3	Combined estimating equation	56
3.3	Poststratification Estimation	64
3.3.1	Estimation of a Total	64
3.3.2	Estimation of Nonlinear Parameters	69
3.4	Analyses of Categorical Complex Survey Data	71
3.4.1	Full-sample Goodness-of-Fit Tests	72
3.4.2	Inverse-sampling Goodness-of-fit Tests	75
3.4.3	Independence Tests for a Two-way Table	81
3.5	Quasi-Score Tests	86

II Analyses of Cluster-Correlated Biological Data Using Inverse Sampling 91

4	Analysis of Cluster-Correlated Data with Nonignorable Cluster Sizes	92
4.1	Introduction	92
4.2	Method of Liang-Zeger	95
4.3	Separate Estimating Equations	100
4.4	Combined Estimating Equation	102
4.5	Mean Estimating Equation	105
4.6	Efficiency of the Mean Estimating Equation Estimator	107
4.6.1	Mean Model	107

4.6.2	Linear regression model	109
4.6.3	Logistic linear regression	110
4.7	Wald tests for Ignorability of Cluster Sizes	111
4.8	Simulation Study	114
4.8.1	Mean Model	114
4.8.2	Linear regression	116
4.8.3	Logistic linear regression	121
4.8.4	Case of small K	124
5	Analyses of Cluster-Correlated Binary Outcomes Data	129
5.1	Introduction	129
5.2	Conditional Logistic Regression	131
5.2.1	Matched-Pair Case-Control Design	133
5.3	Separate Conditional Estimating Equations	134
5.4	Mean Conditional Estimating Equation	136
5.5	Simulation Study	140
6	Analysis of Data With Nonignorable Cluster sizes and Two Types of Correlation	144
6.1	Introduction	144
6.2	A Marginal Model	147
6.2.1	Relative Efficiencies	149
6.3	Total Independence Estimating Equations and Informative Cluster Sizes	159
6.4	Inverse-Sampling Estimating Equations	161

6.4.1	Separate Longitudinal Independence Estimating Equations . .	162
6.4.2	Combined Longitudinal Independence Estimating Equation . .	163
6.4.3	Mean Longitudinal Independence Estimating Equation	167
7	Conclusions and Future Research	170
7.1	Conclusions	170
7.2	Future Research	172
	Appendix I- Splus Syntax to Implement Mean Estimating Equation	174
	Appendix II - Splus Syntax to Implement CLR for Cluster-correlated Matched	
	Case-control Data	175
	References	176

List of Tables

3.1	Actual significance levels (%) of the design-based and inverse-sampling Wald statistics, X_W^2 and $X_{gc,W}^2$ respectively, $\mathbf{p} = (1/K, 1/K, 1/K)'$; $K = 3, \bar{\lambda} = 2, a = 0, m = 10, \alpha = 5\%$	80
3.2	Actual significance levels (%) of $X_c^2, X_S^2, X_{gc,c}, X_{gc,W}^2, K = 3, \bar{\lambda} = 2, a = 0, m = 10, \alpha = 5\%, \mathbf{p} = (1/K, 1/K, 1/K)'$	80
4.1	Size and power of the Wald tests for ignorability of cluster sizes based on 1000 simulations: mean model, nominal level test=0.05, and $K = 100$ clusters	115
4.2	Means and mean squared errors of estimators of μ , relative biases of variance estimators and coverage probabilities of associated 0.95-level normal theory intervals based on 1000 simulations: mean model with $\mu = -0.2$ and $K = 100$ clusters.	117
4.3	Size and power of the Wald tests for ignorability of cluster sizes based on 1000 simulations: linear regression, nominal level test=0.05, and $K = 100$ clusters	119

4.4	Means and mean squared errors of estimators of β_0 and β_1 and coverage probabilities of associated 0.95-level normal theory intervals based on 1000 simulations: linear regression with $\beta_0 = 0.4$, $\beta_1 = -0.2$, $K = 100$ clusters and $g = 1000$ repeated subsampling.	120
4.5	Size and power of the Wald tests for ignorability of cluster sizes based on 1000 simulations: logistic linear regression, nominal level test=0.05, and $K = 100$ clusters	123
4.6	Means and mean squared errors of estimators of β_0 and β_1 and coverage probabilities of associated coverage probabilities of 0.95-level normal theory intervals based on 1000 simulations: logistic linear regression with $\beta_0 = -1.0986$, $\beta_1 = 0.4796$ and $K = 100$ clusters.	127
4.7	Proportion of negative variance estimates(%) under the nested error model (4.8.38) with $K = 12$ clusters and n_i cluster sizes ranging from 1 to 5: 1000 simulation runs, $\beta_0 = 51$, $\beta_1 = 0.329$, $\beta_2 = -0.134$, $\sigma_\alpha^2 = 140$ and $\sigma_\epsilon^2 = 150$	128
5.1	Means and mean squared errors of estimators of β , relative biases of variance estimators and coverage probabilities of associated 0.95 normal theory intervals based on 1000 simulations; $\beta = 0.3$ and $K = 100$ clusters	142
6.1	Percentage relative asymptotic efficiency of $\hat{\beta}_W$ to $\hat{\beta}_Q$, the true correlation matrix is $C_i(\phi, \alpha) = C_{1i}(\phi) \otimes C_{2i}(\alpha)$, $C_{1i}(\phi)$ is exchangeable, $C_{2i}(\alpha)$ is AR(1)	152

6.2	Percentage relative asymptotic efficiency of $\hat{\beta}_W$ to $\hat{\beta}_Q$, the true correlation matrix is $C_i(\phi, \alpha) = C_{1i}(\phi) \otimes C_{2i}(\alpha)$, $C_{1i}(\phi)$ is exchangeable, $C_{2i}(\alpha)$ is MA(1)	153
6.3	Percentage relative asymptotic efficiency of $\hat{\beta}_W$ to $\hat{\beta}_Q$. Use α in $C_{2i}(\alpha) = \text{true MA}(1)$ and α_0 in $C_{2i}^*(\alpha) = \text{working EXC or AR}(1)$. .	156
6.4	Percentage relative asymptotic efficiency of $\hat{\beta}_W$ to $\hat{\beta}_Q$. Use α in $C_{2i}(\alpha) = \text{true AR}(1)$ and α_0 in $C_{2i}^*(\alpha) = \text{working EXC or MA}(1)$. .	158

Chapter 1

Introduction

This thesis consists of two main parts. Part I focuses on the use of inverse-sampling methods to analyze complex survey data. Part II discusses analysis of cluster-correlated biological data using inverse-sampling methods.

It is well known that application of classical statistical methods to analyze complex survey data without taking into account the survey design features can lead to erroneous inferences. In the last two decades many methods have been developed that account for the survey design, but these methods require additional information such as survey weights, design effects or cluster identification for microdata. Most of these methods also require specialized computer software such as SUDAAN and WesVar.

Inverse sampling proposed by Hinkins, Oh and Scheuren (1997) provides an alternative approach by undoing the complex survey data structures so that standard methods can be applied. In this method, repeated subsamples with unconditional simple random sampling structure are drawn and each subsample analyzed by stan-

dard classical methods and the results combined to increase efficiency.

In Part I of this thesis, we discuss the use of the estimating equations approach to inverse sampling to estimate a population total, population ratios, census linear regression and logistic regression parameters. We also use this approach to study post-stratification estimation of a population total and nonlinear parameters, goodness-of-fit tests for categorical complex survey data, and quasi-score tests of census regression parameters. Detailed discussions of these methods are provided in Chapter 2 and Chapter 3.

Part II of the thesis consists of Chapter 4, Chapter 5 and Chapter 6. In Chapter 4, we discuss methods for analyzing cluster-correlated biological (prospective) data. Most of the methods for analyzing data of this type implicitly assume that the cluster sizes are ignorable. When this assumption fails, these methods may lead to asymptotically invalid inferences. We study the performance of the Liang and Zeger (1986) independence estimating equations (IEE) method under both cluster size ignorability and nonignorability. We propose alternative methods, combined estimating equation (CEE) and mean estimating equation (MEE) that remain valid under both cluster size ignorability and informativeness.

Chapter 5 focuses on methods for analyzing cluster-correlated binary outcome data (generally retrospective data, e.g., case-control studies). We discuss the strengths and weaknesses of the commonly used conditional logistic regression method. The method makes the assumption that the within-cluster correlation structure is induced only by the random cluster effects in a cluster-specific mixed effects logistic regression model. Thus it can lead to asymptotically invalid inferences when this assumption

fails. We discuss alternative methods that lead to asymptotically valid inferences when the within-cluster correlation structure is induced by random cluster effects in a cluster-specific mixed effects logistic regression model or by other unmeasured factors.

Chapter 6 focuses on a number of approaches to analyze data with two types of correlation: within-cluster and longitudinal correlations. Most methods for analyzing this type of data implicitly assume that the cluster sizes are ignorable (see Chapter 4 for methods to analyze data with one type of correlation - longitudinal or within-cluster). These methods can lead to asymptotically invalid inferences when the assumption of cluster size ignorability fails. Chapter 6 discusses alternative methods that lead to asymptotically valid inferences under both ignorable and informative cluster sizes.

In Chapter 7 we give a brief summary of the main results of this thesis and some suggestions for future research.

1.1 Results of Chapter 2

1.1.1 Inverse Sampling Algorithms

Suppose s_0 is a sample of observations drawn from a finite population of size N according to a specified complex survey design $p_0(s_0)$. Suppose we wish to draw a subsample s^* of size m from s_0 so that the unconditional probability of s^* , $p(s^*)$ matches simple random sampling with $p(s^*) = 1 / \binom{N}{m}$ either exactly or approximately. Then we

have

$$p(s^*) = \sum_{s_0 \supset s^*} p_0(s_0) p(s^* | s_0), \quad (1.1.1)$$

where $p(s^* | s_0)$ is the conditional probability of choosing s^* . Provided $p(s^* | s_0)$ is independent of s_0 , it follows from (1.1.1) that

$$p(s^* | s_0) = \frac{p(s^*)}{\sum_{s \supset s^*} p_0(s_0)}. \quad (1.1.2)$$

By specifying $p(s^*)$ and $p_0(s_0)$, the inverse sampling algorithms follow from (1.1.2). Subsection 2.2.1 discusses an exact inverse sampling scheme for a stratified simple random sampling using (1.1.2). A number of approximate inverse sampling schemes for one-stage cluster sampling designs are presented in Subsection 2.2.2.

1.1.2 Estimation of Simple Parameters

Suppose θ is a simple census population parameter (e.g, a population total) of interest and s_0 a sample of observations drawn from the population according to some complex survey design. We assume that an inverse sampling algorithm exists to produce subsamples for some simpler design. Let $\hat{\theta}_j^*$ and $\hat{V}_j^*$ denote the estimator and variance estimator produced from the j -th subsample for a sequence of g independent subsamples s_j^* ($j = 1, \dots, g$). Then an inverse-sampling estimator of θ is $\hat{\theta}_g = g^{-1} \sum_{j=1}^g \hat{\theta}_j^*$ and the associated variance estimator

$$\hat{V}(\hat{\theta}_g) = \frac{1}{g} \sum_{j=1}^g \hat{V}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\theta}_j^* - \hat{\theta}_g)^2, \quad (1.1.3)$$

where $\hat{V}_j^*$ is the variance estimator of $\hat{\theta}_j^*$ based on the subsample design. For estimation of a population total, we show that $\hat{\theta}_g$ converges almost surely to θ and $\hat{V}(\hat{\theta}_g)$

converges almost surely to $\widehat{V}(\hat{\theta})$ as $g \rightarrow \infty$, where $\hat{\theta}$ and $\widehat{V}(\hat{\theta})$ are the estimator and the associated variance estimator based on the original sample, s_0 . In addition, we present an asymptotic normality for $\hat{\theta}_g$, see Section 2.4.

Some properties of approximate inverse-sampling methods are presented in Section 2.4.3. For estimating a population total we show that the approximate matching methods for the one-stage or the two-stage cluster sampling designs lead to inverse sampling estimators that are consistent but the associated variance estimators can lead to overestimation if the sampling fraction is not small.

Section 2.6 discusses the generalized regression estimator for estimating a population total using an auxiliary information. We show that an inverse-sampling generalized regression estimator and the associated variance estimator converge almost surely to the corresponding full-sample estimator and the full-sample variance estimator respectively as $g \rightarrow \infty$.

An alternative approach to using an auxiliary information is to consider the inverse sampling design as a special case of a two phase design. Phase one of this design corresponds to the original sample s_0 and phase two consists of the g subsamples. Section 2.5 discusses an inverse-sampling ratio estimator for a population total. We show that the inverse-sampling ratio estimator is more precise for estimating a population total.

In Section 2.7, basic properties of inverse-sampling for estimating quadratic population parameters such as population variances and population covariances are presented.

1.2 Results of Chapter 3

1.2.1 Estimation of Complex Parameters

The inverse-sampling estimation methods discussed in Chapter 2 although useful for simple parameter estimation can lead to asymptotically invalid inferences for complex census population parameters, e.g., census logistic regression parameters. To rectify this limitation, Chapter 3 of the thesis discusses a combined estimating equation approach to inverse sampling. This approach permits asymptotically valid inferences on nonlinear parameters such as census linear regression and logistic regression parameters. Suppose θ denotes a census population parameter (which may be a vector of parameters), we show that under mild regularity conditions, the combined estimating equation estimator, $\hat{\theta}_{gc}$ of θ converges in probability to the full-sample estimator, $\hat{\theta}$; the inverse-sampling variance estimator, $\hat{V}_L(\hat{\theta}_{gc})$ of $\hat{\theta}_{gc}$ also converges in probability to the full-sample variance estimator, $\hat{V}_L(\hat{\theta})$ as the number of repeated subsampling, g goes to infinity. In addition, we show that $\hat{\theta}_{gc}$ is asymptotically normally distributed as $g \rightarrow \infty$ and $n \rightarrow \infty$, where n is the size of the original sample, s_0 .

1.2.2 Some Applications

The remainder of Chapter 3 discusses application of the combined estimating equation approach to poststratification estimation of a population total (see Section 3.3.1) and census nonlinear parameters (see Section 3.3.2).

Section 3.4 studies basic properties of inverse-sampling Wald and Pearson test statistics for goodness-of-fit problems using one-stage and two-stage cluster sampling

designs. Asymptotic normality of the inverse-sampling estimator, $\hat{p}_g$ of a population proportion, p is also presented.

In Section 3.5, inferences on a model parameter θ using inverse-sampling Wald and quasi-score tests are studied. In particular, the focus is on testing a null hypothesis of the form $H_0 : \theta_2 = \theta_{20}$, where θ is partitioned as $\theta = (\theta'_1, \theta'_2)'$.

1.3 Results of Chapter 4

Results of Chapter 4 focus primarily on the strengths and weaknesses of methods for analyzing cluster-correlated biological data with ignorable or informative cluster sizes.

Using a mean model, a nested error linear regression model and a logistic regression model, we show that when cluster sizes are informative, the Liang and Zeger (1986) independence estimating equations (IEE) method lead to bias estimators of the regression parameters, see Section 4.2.

Section 4.3 summarizes a computationally intensive inverse sampling method proposed by Hoffman, Sen and Weinberg (2001). This method leads to asymptotically valid inferences under both ignorable and informative cluster sizes. However, it can produce unstable variance estimators for small samples, see Section 4.8.4.

Properties of a combined estimating equation approach are presented in Section 4.4. This proposed method also leads to asymptotically valid inferences under both ignorable or nonignorable cluster sizes. Even though it is also a repeated subsampling method, it is computationally less intensive and less likely to lead to unstable

variance estimators for small samples, see Section 4.8.4. An alternative method we propose is mean estimating equation, see Section 4.5. This method has properties similar to the combined estimating equation method, but it is not based on resampling. It does not produce unstable variance estimators for small samples. It is also computationally much simpler, since it requires only minor adjustments to standard statistical softwares to implement.

In Section 4.6, we study the efficiency of the mean estimating equation estimator, $\hat{\theta}_M$ of θ compared with the Liang-Zeger IEE estimator, $\hat{\theta}_I$ using a mean model, a nested error linear regression model and a logistic regression model.

Section 4.7 discusses a test for cluster size ignorability by simply comparing the mean estimating equation estimator, $\hat{\theta}_M$ with the Liang-Zeger IEE estimator $\hat{\theta}_I$. Asymptotic normality of the proposed test is presented. Performance of the proposed methods and the proposed cluster size ignorability test are assessed by limited simulation studies. The results of these simulation studies are reported in Section 4.8.

1.4 Results of Chapter 5

Chapter 5 discusses methods to analyze biological cluster-correlated binary outcome (retrospective) data. Section 5.2 summarizes strengths and weaknesses of the commonly used conditional logistic regression method. This method can lead to asymptotically invalid inferences when the within-cluster correlation structure is induced by unmeasured factors other than the random cluster effects.

Rieger and Weinberg (2002) proposed a simple but computationally intensive

method (based on inverse sampling) that is asymptotically valid when the within-cluster correlation structure is induced by random cluster effects or by other unmeasured factors. We summarize their results in Section 5.3.

In section 5.4, we propose an alternative method, mean conditional estimating equation (MCEE), that leads to asymptotically valid inferences when the within-cluster correlation structure is induced solely by the random effects or by other unmeasured factors. Under mild regularity conditions, we present an asymptotic normality theory for the MCEE estimator, $\hat{\theta}_M$, of the regression parameter, θ . To assess the performance of the proposed methods, a limited simulation study was conducted, see Section 5.5.

1.5 Results of Chapter 6

Most methods for analyzing cluster-correlated biological data focuses on data with one type of correlation as discussed in Chapter 4 and Chapter 5. One problem that has received less attention is the analysis of data with two types of correlation: within-cluster correlation and longitudinal correlation. Even so, most of the methods for analyzing this type of data implicitly assume that the cluster sizes are ignorable. When this assumption fails, the resulting inferences may be asymptotically invalid. Chapter 6 studies properties of a number of proposed methods that are asymptotically valid when cluster sizes are ignorable or informative. These methods are also valid for data with outcomes having these two types of correlation.

In section 6.2, we study the efficiency of Liang-Zeger type generalized estimating

equations estimator $\hat{\beta}_W$ of regression parameter β when cluster sizes are ignorable.

A nested error linear regression model is used to illustrate the bias of the Liang-Zeger type total independence estimating equations (TIEE) estimator, $\hat{\beta}_I$ when cluster sizes are nonignorable, see Section 6.3. Here the total independence refers to assumption of working within-cluster and longitudinal independence.

Section 6.4 discusses properties of our proposed methods. The strengths and weaknesses of these methods are also discussed. The method we propose are (i) separate longitudinal independence estimating equations (SLIEE); (ii) combined longitudinal independence estimating equation (CLIEE) and (iii) mean longitudinal estimating equation (MLIEE).

1.6 Conclusions and Future Research

In Chapter 7, we give a summary of the main results of the thesis and some suggestions for future research.

Part I

Analysis of Complex Survey Data

Using Inverse Sampling

Chapter 2

Basic Theory of Inverse Sampling

2.1 Introduction

There is a fairly clear distinction between the focus of traditional sample survey methodology and that of the rest of applied statistics. Survey samplers have concentrated on developing efficient (but complicated) ways of drawing samples to estimate rather simple quantities (population means, proportions, totals, etc.). Most other applied statisticians, by contrast, have concentrated on developing sophisticated methods for fitting very complicated models, but assuming a rather simple sampling structure (often that the observations are independent).

In reality, data from complex surveys are often used to fit complicated models. For example, people may want to use data from a Labour Force Survey, based on stratified multistage sampling, to characterize the association between education and unemployment levels. They might want to use data from large scale health surveys to study the association between housing conditions or poverty and morbidity, and

so on. Extending the range of application of standard methods so that they can be applied to data from complex sample surveys, involving multi-stage sampling and variable selection probabilities, can be difficult and cumbersome; see e.g., Skinner, Holt and Smith (1989).

How do practitioners deal with the complexities of survey data structures such as unequal selection probabilities, clustering and stratification? Adapting a quote from Hinkins, Oh and Scheuren (1997) (abbreviated HOS hereafter): "If your only tool is a hammer, every problem looks like a nail!"; the hammer available to most people is one of the big statistical packages such as SAS, SPSS etc. Most people still just push their data through a standard program and ignore the survey data structures. This is in spite of the fact that a great deal of effort over the last two decades has been spent on developing methods to analyze survey data that take account of design features (see e.g., Skinner, Holt and Smith, 1989), and specialized programs such as SUDAAN or WesVar are now available to implement some of these methods.

An alternative to developing complex new tools is to work backwards: instead of tailoring the methods to fit the data, tailor the data to fit the methods. HOS developed an approach along these lines. Their basic idea is to avoid the pain caused by a complicated sample by choosing a subsample (inverse-sample) that has a simple random sample structure unconditionally (or at least has a structure that is considerably simpler to handle than the original sample). Obviously this involves some loss in efficiency, especially if the subsample is very much smaller than the original sample, as often turns out to be necessary. However, we can increase the efficiency by repeating the process independently many times and averaging the results.

Is it possible to produce subsamples with the desired properties? The answer is often “yes”, although the resulting subsample size, m , might have to be small (in fact, no more than $m = 2$ for some standard stratified multistage designs). HOS give algorithms for producing simple random inverse samples for a number of standard designs.

Section 2.2 summarizes the inverse sampling schemes for a number of standard complex survey designs. These schemes include both exact and approximate methods in terms of matching simple random sampling. In Section 2.3, we study some properties of the inverse sampling algorithms and develop some basic theory of inverse sampling. Section 2.4 illustrates some of the strengths and weaknesses of the inverse sampling method for the special case of estimating linear population parameters (e.g., a population total). We propose an inverse-sampling ratio estimator of a population total in Section 2.5. An inverse-sampling generalized regression estimator of a population total is introduced in Section 2.6. In Section 2.7, we study properties of inverse sampling for estimating quadratic population parameters (e.g., a population variance).

2.2 Inverse Sampling Algorithms

In this section we summarize some inverse sampling schemes, proposed by Hinkins, Oh and Scheuren (1997). These schemes include both exact and approximate methods in terms of matching simple random sampling (SRS) unconditionally.

Suppose we have a sample s_0 of observations drawn from the finite population of

size N according to a specified complex design. We wish to draw a subsample s^* of size m from s_0 such that the unconditional probability of s^* , $p(s^*)$, matches simple random sampling with $p(s^*) = 1 / \binom{N}{m}$, either exactly or approximately. We have

$$p(s^*) = \sum_{s_0 \supset s^*} p_0(s_0) p(s^* | s_0), \quad (2.2.1)$$

where $p_0(s_0)$ is the probability of selecting s_0 and $p(s^* | s_0)$ is the conditional probability of choosing s^* . If $p(s^* | s_0)$ does not depend on s_0 , then it follows from (2.2.1) that

$$p(s^* | s_0) = p_2(s^*) = \frac{p(s^*)}{\sum_{s_0 \supset s^*} p_0(s_0)}. \quad (2.2.2)$$

Denote the first-order and second-order inclusion probabilities corresponding to s^* and s_0 as (π_i^*, π_{il}^*) and (π_i, π_{il}) respectively, where $\pi_i^* = m/N$ and $\pi_{il}^* = m(m-1)/N(N-1)$, $i \neq l$. Similarly, denote the conditional inclusion probabilities as $(\tilde{\pi}_i(s_0), \tilde{\pi}_{il}(s_0))$. If the conditional inclusion probabilities do not depend on s_0 , then we write them as $(\tilde{\pi}_i, \tilde{\pi}_{il})$. It is readily seen that

$$\pi_i^* = \sum_{s_0 \ni i} p_0(s_0) \tilde{\pi}_i(s_0); \quad \pi_{il}^* = \sum_{s_0 \ni i, l} p_0(s_0) \tilde{\pi}_{il}(s_0). \quad (2.2.3)$$

If $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ and $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il}$, then it follows from (2.2.3) that

$$\pi_i^* = \pi_i \tilde{\pi}_i, \quad \pi_{il}^* = \pi_{il} \tilde{\pi}_{il}. \quad (2.2.4)$$

In Section 2.4 we use (2.2.4) to study the properties of inverse sampling for estimating linear population parameters. We also use (2.2.4) in Section 2.7 to study properties of inverse sampling for estimating quadratic population parameters. Note that (π_i^*, π_{il}^*) may correspond to some other simpler sampling design if it is not feasible to match simple random sampling (SRS), e.g., stratified simple random sampling.

2.2.1 Stratified Simple Random Sampling

Suppose that the original sample s_0 is a stratified simple random sample, i.e.,

$$p_0(s_0) = \prod_{h=1}^L \binom{N_h}{n_h}^{-1}, \quad (2.2.5)$$

where $N_h(n_h)$ denotes the number of population (sample) units in stratum h ($= 1, \dots, L$). We wish to draw a subsample s^* of size m such that $p(s^*) = 1 / \binom{N}{m}$, where $N = \sum_{h=1}^L N_h$. Clearly, m cannot be larger than $\min(n_h)$. Let $\mathbf{m} = (m_1, \dots, m_L)'$ denote the (random) number of units in each stratum that belongs to s^* , $0 \leq m_h \leq n_h$, $\sum_{h=1}^L m_h = m$. Noting that the number of terms in $\sum_{s_0 \supset s^*}$ equals $\prod_{h=1}^L \binom{N_h - m_h}{n_h - m_h}$, it follows from (2.2.2) that

$$p(s^*|s_0) = \frac{\prod_{h=1}^L \binom{N_h}{m_h}^{-1}}{\binom{N}{m} \prod_{h=1}^L \binom{n_h}{m_h}}. \quad (2.2.6)$$

The inverse sampling scheme readily follows from (2.2.6): (i) Generate $\mathbf{m}$ from the hypergeometric distribution $f(\mathbf{m}) = \prod_{h=1}^L \binom{N_h}{m_h} / \binom{N}{m}$; (ii) Draw a simple random sample of size m_h , without replacement, from the n_h sample units in stratum h , independently across strata h ($= 1, \dots, L$). HOS specify $p(s^*|s_0)$ first and then verify that it gives $p(s^*) = \binom{N}{m}^{-1}$. Our approach provides the inverse sampling scheme from the specification of $p_0(s_0)$ and $p(s^*)$.

2.2.2 One-Stage Cluster Sampling

HOS studied the case of one-stage cluster sampling in detail. Three sampling designs for s_0 were investigated: (1) Equal cluster sizes, M , and simple random sampling of clusters; (2) Unequal cluster sizes, M_i , and simple random sampling of clusters; (3) Unequal cluster sizes, M_i , and clusters sampled with probability proportional to size (PPS) M_i and with replacement.

Case 1: Equal Cluster Sizes, M , Clusters Sampled by SRS

Suppose s_0 contains k clusters drawn from K clusters in the population $N = KM$ with the probability of selecting s_0 given by

$$p_0(s_0) = \binom{K}{k}^{-1}. \quad (2.2.7)$$

We wish to draw an inverse-sample s^* of size m such that $p(s^*) = 1/\binom{KM}{m}^{-1}$. We observe that m cannot be larger than k . Let $\mathbf{m} = (m_1, \dots, m_k)'$ denote the (random) number of units in each sample cluster that belongs to s^* , $0 \leq m_i \leq m$, $\sum_{i=1}^k m_i = m$. We also note that the number of terms in $\sum_{s_0 \supset s^*}$ equals $\binom{K-m}{k-m}$, it follows from (2.2.2) that

$$p(s^*|s_0) = \frac{1}{\binom{KM}{m}} \frac{\binom{K}{k}}{\binom{K-m}{k-m}}. \quad (2.2.8)$$

The form of $p(s^*|s_0)$ provided by HOS is given by

$$p(s^*|s_0) = \frac{\prod_{i=1}^m \binom{M}{m_i} \binom{K}{k}}{\binom{KM}{m} \binom{K-m}{k-m} \prod_{i=1}^m \binom{M}{m_i}} \cdot 1. \quad (2.2.9)$$

Clearly, the expression $p(s^*|s_0)$ provided by HOS in (2.2.9) and the one we provided in (2.2.8) are essentially the same. The inverse sampling scheme does not follow readily from either (2.2.8) or (2.2.9). In fact, exact matching with SRS is difficult to implement for this design. A partial solution is to use an approximate inverse sampling scheme. This simple approximate scheme selects one element at random from each sample cluster so that the size of s^* is k . We study the properties of this approximate scheme for estimating a finite population total in Section 2.4. HOS used systematic sampling to select one case from each sample cluster, a different approximate inverse sampling scheme.

Case 2: Unequal Cluster Sizes, M_i , Clusters Sampled by SRS

Suppose s_0 contains k clusters drawn from K clusters in the population $N = \sum_{i=1}^K M_i$. We wish to draw a subsample sample s^* of fixed size m . Exact or approximate inverse sampling scheme to obtain SRS subsample of fixed size m does not exist. HOS proposed an alternative method that artificially enlarges the population to equal cluster size case and then applies subsampling used in Case 1. We first force all clusters to have the same size by adding an appropriate number of pseudo-units to bring them up to the size of the largest sample cluster. Then we take one unit at random from each sample cluster, and discard any pseudo-units to obtain the final sample. This

approximate method makes $p(s^*|s_0)$ depend on s_0 because the conditional probability depends on $M(s_0)$, the size of the largest sample cluster.

Case 3: Unequal Cluster Sizes, M_i , Clusters Sampled by PPS

For the case of probability proportional to size (PPS) sampling with replacement of unequal size clusters, HOS proposed a simple method of subsampling which gives $p(s^*) = (1/N)^k$, where s^* now denotes an ordered simple random sample with replacement selected from the $N = \sum_{i=1}^K M_i$, units in the population. Viewing the sample clusters as ordered, we select one unit at random from each sample cluster. Note that the same cluster might appear more than once in the ordered sample. Denote the size of the cluster drawn in the i -th PPS draw by M'_i , then

$$p(s^*) = \left(\prod_{i=1}^k \frac{M'_i}{N} \right) \left(\prod_{i=1}^k \frac{1}{M'_i} \right) = \left(\frac{1}{N} \right)^k, \quad (2.2.10)$$

where $\prod_{i=1}^k (M'_i/N)$ is the probability of drawing the ordered cluster sample. Note that s_0 is the ordered PPS sample and we have only one term in the summation in (2.2.1).

If the clusters are drawn with inclusion probabilities $\pi_i = kM_i/N$ and without replacement, then it is not possible to match SRS. However, we can treat the clusters as if they were drawn with replacement, as done in practice, and then apply the scheme for Case 3. This will lead to overestimation of variance, but the overestimation is not serious if the sampling fraction k/K is small (see Section 2.4).

2.2.3 Two-Stage Cluster Sampling

HOS also studied two-stage sampling for the following cases: (1) Equal cluster sizes, M , and k clusters selected by simple random sampling in the first stage; simple random subsample of equal size, m , drawn independently within each sampled cluster (PSU). (2) Unequal cluster sizes, M_i , and k clusters sampled with PPS and with replacement; simple random subsamples of unequal sizes, m_i , drawn independently within each cluster in the with replacement sample.

Case 1: Equal Cluster Sizes, SRS at both Stages

As in the case of one-stage cluster sampling, exact method of inverse sampling is difficult to implement. A simple approximate method of inverse sampling selects one unit at random from each of the k subsamples.

Case 2: Unequal Cluster Sizes, PPS at Stage one and SRS at Stage two

As in Case 3 of the uni-stage cluster sampling, we simply select one unit at random from each of the ordered subsamples. HOS suggested a different method: Take a simple random sample with replacement of k clusters first and then with each selected cluster take one unit at random from the corresponding subsample. It appears that the first stage inverse sampling of clusters is not necessary. To see this, we note that

$$p_0(s_0) = \prod_{i=1}^k \left\{ \left(\frac{M'_i}{N} \right) \frac{1}{\binom{M'_i}{m'_i}} \right\},$$

where m'_i is the subsample size associated with the cluster selected in the i -th draw ($i =$

$1, \dots, k$). We wish to draw a subsample s^* of size k such that $p(s^*) = (1/N)^k$, where $N = \sum_{i=1}^K M_i$. Also the number of terms in $\sum_{s_0 \supset s^*}$ equals $\prod_{i=1}^k \binom{M'_i - 1}{m'_i - 1}$ and

$$\sum_{s_0 \supset s^*} p_0(s_0) = \prod_{i=1}^k \left\{ \left(\frac{M'_i}{N} \right) \frac{\binom{M'_i - 1}{m'_i - 1}}{\binom{M'_i}{m'_i}} \right\} = \prod_{i=1}^k \frac{m'_i}{N}.$$

It follows from (2.2.2) that $p(s^*|s_0) = \prod_{i=1}^k (1/m'_i)$ and hence the subsampling scheme readily follows.

2.2.4 Stratified Two-Stage Sampling

Suppose we have a two-stage sample from each stratum, where the clusters are sampled with PPS with replacement and subsampling done independently within each sampled cluster. Using the inverse sampling procedure of Case 2, Subsection 2.2.3, we get simple random samples from each stratum. We can then apply the method of Subsection 2.2.1, treating the inverse samples as if drawn without replacement to get an inverse sample of size $k_0 = \min_h(k_h)$, where k_h is the number of sampled clusters in stratum h . In the important case of $k_h = 2$ psu's sampled from each stratum, the inverse sample size, k_0 , is only two.

2.3 Basic Properties of Inverse Sampling

Suppose that we are interested in estimating some population parameter, θ , and we have a sample, s_0 , of observations drawn from the population according to some complex design, $p_0(s_0)$. We assume that we have an inverse sampling algorithm that can produce samples from some simpler design. This design will often be simple random sampling, but we can extend the range of applications considerably by allowing for the possibility of more general (sub-)designs; for example, stratified SRS when the original sample is a stratified two-stage sample. Our only requirement for the simpler design is that we can produce an estimator of the quantity of interest, θ , together with an estimator of its variance. Let $\hat{\theta}_j^*$ and $\hat{V}_j^*$ denote the estimator and variance estimator produced from the j -th subsample when we generate a sequence of g independent subsamples s_j^* ($j = 1, \dots, g$). Note that the $\hat{\theta}_j^*$'s are not unconditionally independent when averaged over the distribution of the initial sample, s_0 . An estimator of θ based on the g subsamples is

$$\hat{\theta}_g = \frac{1}{g} \sum_{j=1}^g \hat{\theta}_j^*. \quad (2.3.11)$$

We denote the estimator based on the full sample s_0 as $\hat{\theta}$.

Theorem 2.1 1. *Conditional on the original sample, s_0 , $\hat{\theta}_g$ converges almost surely*

to $E(\hat{\theta}_1^|s_0) = \hat{\theta}_\infty$, say, as $g \rightarrow \infty$.*

2. $E(\hat{\theta}_g) = E(\hat{\theta}_1^*)$.

3. $Var(\hat{\theta}_g) = Var(\hat{\theta}_\infty) + \frac{1}{g} E\{Var(\hat{\theta}_1^*|s_0)\}$.

4. If $r_g = \frac{\text{Var}(\hat{\theta}_g)}{\text{Var}(\hat{\theta}_\infty)}$, then $r_g = 1 + \frac{r_1-1}{g}$.

Proof

Result 1 of Theorem 2.1 follows directly from (2.3.11) by noting that conditional on s_0 , the subsampling estimators, $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$ are independent identically distributed bounded random variables.

Result 2 of Theorem 2.1 follows from the standard relationship between conditional and unconditional expectations:

$$E(\hat{\theta}_g) = E\{E(\hat{\theta}_g|s_0)\} = E\{E(\hat{\theta}_1^*|s_0)\} = E(\hat{\theta}_1^*).$$

Result 3 of Theorem 2.1 follows from the corresponding result for variances, and the conditional independence of the $\hat{\theta}_j^*$'s given s_0 :

$$\text{Var}(\hat{\theta}_g) = \text{Var}\{E(\hat{\theta}_g|s_0)\} + E\{\text{Var}(\hat{\theta}_g|s_0)\} = \text{Var}(\hat{\theta}_\infty) + \frac{1}{g}E\{\text{Var}(\hat{\theta}_1^*|s_0)\}.$$

Result 4 of Theorem 2.1 follows directly from Result 3 of Theorem 2.1. ■

Result 4 of Theorem 2.1 demonstrates that increasing the number of subsamples, g , does indeed increase the efficiency of $\hat{\theta}_g$. More precisely, the variance ratio r_g has the form $a + b/g$. If the subsample estimator, $\hat{\theta}_1^*$, is unbiased for θ , then so is the inverse-sampling estimator, $\hat{\theta}_g$. However, if $\hat{\theta}_1^*$ has bias of order m^{-1} , where m denotes the subsample size, then $\hat{\theta}_g$ has exactly the same bias. Since m will usually be very much smaller than the original sample size, this bias can be appreciable. This is a serious limitation of $\hat{\theta}_g$ in the nonlinear cases, such as ratios and regression

coefficients. In Section 3.1, we propose an alternative estimator of θ based on the estimating equations (EE) approach. This estimator is asymptotically unbiased for any m as the size of s_0 increases, unlike $\hat{\theta}_g$.

Result 4 of Theorem 2.1 can be used to determine the number of subsamples, g needed to obtain reasonable efficiency. For example, HOS give an example in which $r_1 = 29.3$. The original sample was a very efficient stratified random sample with $n = 15,618$ observations taken from the Statistics of Income corporate survey, while the subsample was a simple random sample of $m = 2,224$ observations. A single subsample is relatively inefficient. However, in this case, repeated inverse sampling recovers all the information in the original sample in the limit. Applying Result 4 of Theorem 2.1 leads immediately to the following table:

g	1	10	100	1000
r_g	29.3	3.83	1.28	1.03

(HOS produced these same results by simulation but this is unnecessary in view of Result 4 of Theorem 2.1.) We see that $g = 100$ subsamples would be adequate for many purposes and that we obtain almost full efficiency with $g = 1000$.

The fact that $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$ are not unconditionally independent means that estimating $Var(\hat{\theta}_g)$ is not completely straightforward. However, a relatively simple variance estimator may be obtained using Theorem 2.2 below.

Theorem 2.2 *The variance of $\hat{\theta}_g$ is given by*

$$Var(\hat{\theta}_g) = Var(\hat{\theta}_1^*) - \frac{g-1}{g} E\{Var(\hat{\theta}_1^*|s_0)\}. \quad (2.3.12)$$

Proof

Theorem 2.2 follows from applying Results 3 of Theorem 2.1 with $g = 1$ to obtain

$$\text{Var}(\hat{\theta}_\infty) = \text{Var}(\hat{\theta}_1^*) - E\{\text{Var}(\hat{\theta}_1^*|s_0)\},$$

and then substituting this expression for $\text{Var}(\hat{\theta}_\infty)$ in Result 3 of Theorem 2.1 for general g . ■

We can estimate the first term of (2.3.12) by $\hat{V}_j^*$ for $j = 1, \dots, g$, and hence by their average $g^{-1} \sum_{j=1}^g \hat{V}_j^*$. In addition, the quantity

$$s_{\theta_g}^2 = \frac{1}{g-1} \sum_{j=1}^g (\hat{\theta}_j^* - \hat{\theta}_g)^2$$

gives an unbiased estimator of $E\{\text{Var}(\hat{\theta}_1^*|s_0)\}$ since $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$ are conditionally independent given the initial sample, s_0 . This leads to an estimator of $\text{Var}(\hat{\theta}_g)$ of the form

$$\hat{V}_g = \frac{1}{g} \sum_{j=1}^g \hat{V}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\theta}_j^* - \hat{\theta}_g)^2. \quad (2.3.13)$$

Properties of the variance estimator $\hat{V}_g$ depend on the properties of the subsample estimator $\hat{V}_j^*$. For example, if $\hat{V}_j^*$ is unbiased, then $\hat{V}_g$ is also unbiased.

For the special case of population total $\theta = Y$ and simple random subsampling, i.e., $p(s^*) = 1/\binom{N}{m}$, we have $\hat{\theta}_j^* = \hat{Y}_j^* = N\bar{y}_j^*$ and $\hat{Y}_j^*$ is unbiased for Y with unbiased variance estimator $\hat{V}_j^* = N^2(m^{-1} - N^{-1})s_{jy}^{*2}$, where $\bar{y}_j^*$ is the mean and s_{jy}^{*2} is the variance of the j -th subsample. The variance estimator $\hat{V}_g$ of $\hat{\theta}_g = \hat{Y}_g =$

$g^{-1} \sum_{j=1}^g (N\bar{y}_j^*)$, given by (2.3.13), is unbiased, and it reduces to

$$\hat{V}_g = \frac{1}{g} \sum_{j=1}^g \hat{V}_j^* - \frac{N^2}{g} \sum_{j=1}^g (\bar{y}_j^* - \bar{y}_g)^2, \quad (2.3.14)$$

where $\bar{y}_g = g^{-1} \sum_{j=1}^g \bar{y}_j^*$.

2.4 Estimating Linear Population Parameters

In this section we study the properties of inverse-sampling methods for estimating linear population parameters (e.g., population total), when exact or approximate matching methods can be implemented on the original design. Properties of the exact matching methods are studied in Subsection 2.4.1 and Subsection 2.4.2. In Subsection 2.4.3 we study the properties of the approximate matching methods.

2.4.1 Exact Matching

We show here that for estimating linear population parameters, when exact matching methods can be implemented, and provided certain mild conditions are satisfied, the inverse-sampling estimator $\hat{\theta}_g$ converges to the full-sample estimator, $\hat{\theta}$, and also the inverse-sampling variance estimator of $\hat{\theta}_g$ converges to the full-sample variance estimator of $\hat{\theta}$ as $g \rightarrow \infty$.

Consider the estimation of a linear population parameter given by

$$T_1 = \sum_{i=1}^N a_i y_i, \quad (2.4.15)$$

where y_i is the characteristic of a population, N is the size of the population, and a_i is a known quantity for all units in the population. The linear population parameter

reduces to a finite population total when $a_i = 1$ for all units in the population, i.e., $T_1 = Y = \sum_{i=1}^N y_i$. When $a_i = 1$ for all units in the j -th subpopulation (domain, ${}_jD$), then the linear population parameter reduces to $T_1 = {}_jY = \sum_{i \in {}_jD} y_i = \sum_{i=1}^{{}_jN} y_i$, where ${}_jN$ is the size of the j -th domain. The Horvitz-Thompson (H-T) unbiased estimator of T_1 is given by,

$$\hat{T}_1 = \sum_{i \in s_0} \frac{a_i y_i}{\pi_i}, \quad (2.4.16)$$

based on the full sample. Theorem 2.3 establishes conditions under which the corresponding inverse-sampling estimator

$$\hat{T}_{1g} = \frac{1}{g} \sum_{j=1}^g \hat{T}_{1j}^* = \frac{1}{g} \sum_{j=1}^g \sum_{i \in s_j^*} \frac{a_i y_i}{\pi_i^*}, \quad (2.4.17)$$

converges to the full-sample Horvitz-Thompson estimator, $\hat{T}_1$, as $g \rightarrow \infty$.

Theorem 2.3 *Let $\tilde{\pi}_i(s_0)$ be the conditional probability that the i -th unit is selected in the inverse-sample for a given initial sample, s_0 . Suppose that $\hat{\theta}_j^* = \hat{T}_{1j}^*$ is the H-T estimator of a finite population linear parameter $\theta = T_1$ for the j -th subsample. Then the limiting inverse-sampling estimator, $\hat{\theta}_\infty^* = \hat{T}_{1,\infty}$, will be the full-sample H-T estimator, $\hat{T}_1$, if and only if the conditional inclusion probabilities $\tilde{\pi}_i(s_0)$ are constant for all s_0 containing the i -th unit, i.e., $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ for all $s_0 \supset i$.*

Proof

We have

$$\hat{T}_{1j}^* = \sum_{i \in s_j^*} \frac{a_i y_i}{\pi_i^*} = \sum_{i \in s_0} \frac{a_i y_i I_{ij}^*(s_0)}{\pi_i^*},$$

where $I_{ij}^*(s_0)$ takes the value 1 if the i -th unit is included in the j -th inverse-sample s_j^* and 0 otherwise, and π_i^* is the corresponding (unconditional) selection probability.

Thus

$$\hat{T}_{1,\infty} = E(\hat{T}_{11}^* | s_0) = \sum_{i \in s_0} \frac{a_i y_i \tilde{\pi}_i(s_0)}{\pi_i^*}.$$

This is equal to the full-sample H-T estimator $\hat{T}_1 = \sum_{i \in s_0} (a_i y_i / \pi_i)$, if and only if $\tilde{\pi}_i(s_0) = \tilde{\pi}_i = \pi_i^* / \pi_i$.

■

The condition $\tilde{\pi}_i(s_0) = \tilde{\pi}_i$ is a fairly natural one for most sampling designs for which the H-T estimator is used. For the estimation of a total $T_1 = Y$, if the subsamples are all simple random samples of fixed size m , then the estimator for a subsample is simply $N\bar{y}_j^*$, which is the natural estimator under simple random sampling.

Theorem 2.4 establishes conditions under which the inverse-sampling variance estimator, $\hat{V}_{g,HT}$, of $\hat{T}_{1g}$ converges to $\hat{V}_{HT}$, the H-T variance estimator of $\hat{T}_1$ for the original design, as $g \rightarrow \infty$, where

$$\hat{V}_{HT} = \sum_{i,l \in s_0} \sum \frac{\pi_{il} - \pi_i \pi_l}{\pi_i \pi_l \pi_{il}} (a_i y_i)(a_l y_l) \quad (2.4.18)$$

(using result from Cochran, 1977, page 261) and

$$\hat{V}_{g,HT} = \frac{1}{g} \sum_{j=1}^g \hat{V}_{j,HT}^* - \frac{1}{g} \sum_{j=1}^g (\hat{T}_{1j}^* - \hat{T}_{1g})^2$$

with

$$\hat{V}_{j,HT}^* = \sum_{i,l \in s_j^*} \sum \frac{\pi_{il}^* - \pi_i^* \pi_l^*}{\pi_i^* \pi_l^* \pi_{il}^*} (a_i y_i)(a_l y_l). \quad (2.4.19)$$

Note that $\hat{V}_{j,HT}^*$ is the H-T variance estimator of $\hat{T}_{1j}^*$, and $\pi_{ii}^* = \pi_i^*$, $\pi_{ii} = \pi_i$.

Theorem 2.4 *If $\widehat{V}_{j,HT}^*$ is the Horvitz-Thompson (H-T) variance estimator of $\widehat{T}_{1j}^*$ for the j -th subsample, then conditional on s_0 , $\widehat{V}_{g,HT}$ converges to the Horvitz-Thompson (H-T) variance estimator of $\widehat{T}_1$ for the original design, as $g \rightarrow \infty$, if the conditional joint inclusion probabilities are constant for all s_0 containing a given pair (i, l) of units, i.e. $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il}$ for all $s_0 \supset \{i, l\}$.*

Proof

Conditional on s_0 , it follows from (2.3.13) that $\widehat{V}_{g,HT}$ converges almost surely to

$$\widehat{V}_{\infty,HT} = E(\widehat{V}_{1,HT}^* | s_0) - Var(\widehat{T}_{11}^* | s_0) \quad (2.4.20)$$

as $g \rightarrow \infty$. Now, noting that $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il} = \pi_{il}^* / \pi_{il}$, we get

$$\begin{aligned} E(\widehat{V}_{1,HT}^* | s_0) &= \sum_{i,l \in s_0} \sum \frac{\pi_{il}^* - \pi_i^* \pi_l^*}{\pi_i^* \pi_l^* \pi_{il}} \tilde{\pi}_{il} (a_i y_i)(a_l y_l) \\ &= \sum_{i,l \in s_0} \sum \left(\frac{\tilde{\pi}_{il}}{\pi_i^* \pi_l^*} - \frac{1}{\pi_{il}} \right) (a_i y_i)(a_l y_l). \end{aligned} \quad (2.4.21)$$

Further,

$$\begin{aligned} Var(\widehat{T}_{11}^* | s_0) &= \sum_{i,l \in s_0} \sum (\tilde{\pi}_{il} - \tilde{\pi}_i \tilde{\pi}_l) \frac{a_i y_i}{\pi_i^*} \frac{a_l y_l}{\pi_l^*} \\ &= \sum_{i,l \in s_0} \sum \left(\frac{\tilde{\pi}_{il}}{\pi_i^* \pi_l^*} - \frac{1}{\pi_i \pi_l} \right) (a_i y_i)(a_l y_l). \end{aligned} \quad (2.4.22)$$

It now follows from (2.4.20) - (2.4.22) that $\widehat{V}_{\infty,HT} = \widehat{V}_{HT}$. ■

A special case of the above results, is the estimation of a finite population total, Y . This means that by Theorem 2.3 the H-T inverse-sampling estimator of Y converges to the full-sample H-T estimator for large g . Similarly, by Theorem 2.4 the H-T inverse-sampling variance estimator converges to the full-sample H-T variance estimator.

We study an exception to the above variance estimation result given in Theorem 2.4. Consider for example, the Sen-Yates-Grundy(S-Y-G) variance estimator, $\hat{V}_{SYG}$, which is often preferred over the H-T variance estimator, $\hat{V}_{HT}$, because it is more stable and there are several designs for which it is known that it is always non-negative, while $\hat{V}_{HT}$ frequently takes negative values (Cochran, 1977, page 261). The S-Y-G full-sample variance estimator, $\hat{V}_{SYG}$, of $\hat{Y}$ is given by

$$\hat{V}_{SYG} = \sum_{i < l \in s_0} \frac{\pi_i \pi_l - \pi_{il}}{\pi_{il}} \left(\frac{y_i}{\pi_i} - \frac{y_l}{\pi_l} \right)^2. \quad (2.4.23)$$

The S-Y-G inverse-sampling variance estimator, $\hat{V}_{g,SYG}$, of $\hat{T}_{1g} = \hat{Y}_g = g^{-1} \sum_{j=1}^g Y_j^*$ is given by

$$\hat{V}_{g,SYG} = \frac{1}{g} \sum_{j=1}^g \hat{V}_{j,SYG}^* - \frac{1}{g} \sum_{j=1}^g (\hat{Y}_j^* - \hat{Y}_g)^2, \quad (2.4.24)$$

where

$$\hat{V}_{j,SYG}^* = \sum_{i < l \in s_j^*} \frac{\pi_i^* \pi_l^* - \pi_{il}^*}{\pi_{il}^*} \left(\frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*} \right)^2. \quad (2.4.25)$$

Theorem 2.5 below shows that $\hat{V}_{g,SYG}$ does not converge to the S-Y-G variance estimator as $g \rightarrow \infty$, i.e., $\hat{V}_{\infty,SYG} \neq \hat{V}_{SYG}$. If the subsample is a simple random sample unconditionally, i.e., $\pi_i^* = m/N$ and $\pi_{il}^* = m(m-1)/[N(N-1)]$; $i \neq l$, then $\hat{V}_{j,HT}^* = \hat{V}_{j,SYG}^*$ and $\hat{V}_{\infty,SYG} = \hat{V}_{\infty,HT} = \hat{V}_{HT}$, the H-T variance estimator of $\hat{Y}$.

Theorem 2.5 *The inverse-sampling variance estimator (2.4.24) does not converge to the S-Y-G variance estimator (2.4.23) as $g \rightarrow \infty$.*

Proof

Conditional on s_0 , it follows from (2.3.13) that

$$\hat{V}_{\infty, SYG} = E(\hat{V}_{1, SYG}^* | s_0) - Var(\hat{Y}_1^* | s_0) \quad (2.4.26)$$

where

$$Var(\hat{Y}_1^* | s_0) = \sum_{i < l \in s_0} (\tilde{\pi}_i \tilde{\pi}_l - \tilde{\pi}_{il}) \left(\frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*} \right)^2, \quad (2.4.27)$$

provided the subsample size is also fixed (Cochran, 1977, page 260). Further,

$$E(\hat{V}_{1, SYG}^* | s_0) = \sum_{i < l \in s_0} \frac{(\pi_i^* \pi_l^* - \pi_{il}^*)}{\pi_{il}} \left(\frac{y_i}{\pi_i^*} - \frac{y_l}{\pi_l^*} \right)^2. \quad (2.4.28)$$

It now follows that

$$\hat{V}_{\infty, SYG} = \sum_{i < l \in s_0} \frac{\pi_i \pi_l - \pi_{il}}{\pi_{il}} \tilde{\pi}_i \tilde{\pi}_l \left(\frac{y_i}{\pi_i \tilde{\pi}_i} - \frac{y_l}{\pi_l \tilde{\pi}_l} \right)^2. \quad (2.4.29)$$

Comparing (2.4.29) and (2.4.23) we see that $\hat{V}_{\infty, SYG} \neq \hat{V}_{SYG}$.

■

2.4.2 Estimation of a Population Total Using PPS

(i) Unistage Cluster Sampling

As shown in Subsection 2.2.2, exact matching with SRS with replacement exists for probability proportional to size (PPS) sampling with replacement of clusters with unequal sizes M_i . The estimator of a finite population total, Y under PPS is given by

$$\hat{Y}_{pps} = \frac{N}{k} \sum_{i=1}^k \bar{Y}'_i, \quad (2.4.30)$$

where N is the total number of population elements and $\bar{Y}'_i$ is the mean of the cluster selected on the i -th draw. We note that the estimator $\hat{Y}_{pps}$ is not equal to the H-T estimator of Y . The variance estimator of $\hat{Y}_{pps}$ is given by

$$\hat{V}_{pps} = \frac{N^2}{k} \frac{1}{k-1} \sum_{i=1}^k \left(\bar{Y}'_i - \frac{1}{k} \sum_{i=1}^k \bar{Y}'_i \right)^2. \quad (2.4.31)$$

The inverse-sampling estimator corresponding to $\hat{Y}_{pps}$ is given by

$$\hat{Y}_g = \frac{1}{g} \sum_{j=1}^g \hat{Y}_j^*, \quad (2.4.32)$$

where $\hat{Y}_j^* = (N/k) \sum_{i \in s_j^*} y_i$ denotes the estimator of Y from the j -th subsample. The inverse-sampling variance estimator is given by

$$\hat{V}_g = \hat{V}(\hat{Y}_g) = \frac{1}{g} \sum_{j=1}^g \hat{V}(\hat{Y}_j^*) - \frac{1}{g} \sum_{j=1}^g (\hat{Y}_j^* - \hat{Y}_g)^2, \quad (2.4.33)$$

where $\hat{V}(\hat{Y}_j^*) = \{N^2/[k(k-1)]\} \sum_{i \in s_j^*} (y_i^* - \bar{y}_j^*)^2$.

Theorem 2.6 (i) *The inverse-sampling estimator (2.4.32) converges to the full-sample (PPS) estimator (2.4.30); (ii) the inverse-sampling variance estimator (2.4.33) converges to the full-sample (PPS) variance estimator (2.4.31) as $g \rightarrow \infty$.*

Proof of (i) and (ii) of Theorem 2.6

(i) Since $\hat{Y}_1^*, \dots, \hat{Y}_g^*$ are independent identically distributed bounded random variables, we have

$$\hat{Y}_\infty = E(\hat{Y}_1^* | s_0) = \frac{N}{k} \sum_{i=1}^k \bar{Y}'_i = \hat{Y}_{pps},$$

where $\bar{Y}'_i = 1/M'_i \sum_{j=1}^{M'_i} y_{ij}$.

(ii) Now

$$\begin{aligned}
\hat{V}(\hat{Y}_\infty) &= E \left\{ \hat{V}(\hat{Y}_1^*) | s_0 \right\} - V(\hat{Y}_1^* | s_0) \\
&= E \left\{ \hat{V}(\hat{Y}_1^*) | s_0 \right\} - E(\hat{Y}_1^{*2} | s_0) + E^2(\hat{Y}_1^* | s_0) \\
&= E \left\{ \frac{N^2}{k(k-1)} \sum_{i=1}^k (y_i^* - \bar{Y}_1^*)^2 | s_0 \right\} - E \left[\left\{ \frac{N}{k} \sum_{i=1}^k y_i^* \right\}^2 | s_0 \right] + E^2(\hat{Y}_1^* | s_0) \\
&= \frac{N^2}{k(k-1)} \left\{ \sum_{i=1}^k \bar{Y}_i'^2 - \left(\sum_{i=1}^k \bar{Y}_i' \right)^2 \right\} + \hat{Y}_{pps}^2 \\
&= \frac{N^2}{k(k-1)} \sum_{i=1}^k \left(\bar{Y}_i' - \frac{1}{k} \sum_{i=1}^k \bar{Y}_i' \right)^2 \\
&= \hat{V}_{pps}.
\end{aligned}$$

■

This shows that inverse sampling preserves both the full-sample estimator and its variance estimator as $g \rightarrow \infty$.

(ii) Two-Stage Cluster Sampling

Turning to the case of unequal cluster sizes, M_i , we select the clusters with PPS and with replacement, and then draw simple random subsampling of equal size, m , independently within each cluster in the with-replacement sample. The estimator of Y is $\hat{Y}_{pps} = (N/k) \sum_{i=1}^k \bar{y}_i'$ where $\bar{y}_i'$ is the sample mean of the cluster selected in the i -th draw. The variance estimator of $\hat{Y}_{pps}$ is given by

$$\hat{V}_{pps} = \frac{N^2}{k} \frac{1}{k-1} \sum_{i=1}^k \left(\bar{y}_i' - \frac{1}{k} \sum_{i=1}^k \bar{y}_i' \right)^2.$$

The inverse-sampling estimator is given by $\hat{Y}_g = g^{-1} \sum_{j=1}^g \hat{Y}_j^*$, where $\hat{Y}_1^* = (N/k) \sum_{i=1}^k y_i'$, and y_i' is defined as above. It is easy to verify that (i) $\hat{Y}_\infty = \hat{Y}_{pps}$ and (ii) $\hat{V}_\infty = \hat{V}_{pps}$

following similar arguments used under the unistage cluster sampling case. Thus, inverse-sampling preserves both the estimator and the variance estimator.

Theorem 2.7 below is a central limit theorem of the inverse-sampling estimator, $\bar{y}_g$ and Theorem 2.8 gives conditions under which the inverse-sampling variance estimator, $\hat{V}(\bar{y}_g)$ is a consistent estimator of $V(\bar{y})$. The usual formulation of asymptotic theory (including the central limit theorem) for finite populations is to embed the finite population in a sequence of populations, $\{U_\nu\}$ of size N_ν , indexed by ν , where $n_\nu \rightarrow \infty$, $N_\nu \rightarrow \infty$ as $\nu \rightarrow \infty$. We note that n_ν is the size of a complex survey sample, s_ν drawn from U_ν with a sampling design $p_\nu(s_\nu)$. In this thesis, we simplify notation by letting $n \rightarrow \infty$ to mean $n_\nu \rightarrow \infty$, $N_\nu \rightarrow \infty$ as $\nu \rightarrow \infty$. We consider the following regularity conditions:

- C1. $nV(\bar{y}) \rightarrow \sigma^2$ as $n \rightarrow \infty$
- C2. $n \{ \hat{V}(\bar{y}) - V(\bar{y}) \} = o_p(1)$ as $n \rightarrow \infty$
- C3. $\sigma_*^2 = V(\bar{y}_1^* | s_0) = V(\frac{1}{m} \sum_{i \in s_1^*} y_i | s_0)$ is finite
- C4. $n^{\frac{1}{2}}(\bar{y} - \bar{Y}) \rightarrow_d N(0, \sigma^2)$ as $n \rightarrow \infty$

Conditions C1, C2 and C4 are standard assumptions to establish asymptotically valid inferences under complex survey design. Condition C3 is required to establish the boundedness of $(\bar{y}_g - \bar{y})$.

Theorem 2.7 *Let $\tilde{\pi}_i(s_0)$ be the conditional inclusion probability that the i -th unit is selected in the inverse-sample for a given initial sample, s_0 . Suppose $\bar{y} = N^{-1}\hat{Y}$ and $\bar{y}_j^* = m^{-1} \sum_{i \in s_j^*} y_i$ are the full-sample and the j -th inverse-sample estimators*

of $\bar{Y} = N^{-1} \sum_{i=1}^N y_i$, where $\hat{Y}$ is the Horvitz-Thompson estimator of Y , and $\bar{y}_g = g^{-1} \sum_{j=1}^g \bar{y}_j^*$ is the repeated subsampling estimator. Then under conditions C1 - C4, $n^{\frac{1}{2}}(\bar{y}_g - \bar{Y}) \rightarrow_d N(0, \sigma^2)$ as $g \rightarrow \infty$ then $n \rightarrow \infty$ if and only if $\bar{\pi}_i(s_0)$ are constant for all s_0 containing the i -th unit, i.e., $\bar{\pi}_i(s_0) = \bar{\pi}_i$ for all $s_0 \supset i$.

Proof

We first write

$$\sqrt{n}(\bar{y}_g - \bar{Y}) = \frac{\sqrt{n}}{\sqrt{g}} \sqrt{g}(\bar{y}_g - \bar{y}) + \sqrt{n}(\bar{y} - \bar{Y}). \quad (2.4.34)$$

It is sufficient to show that

$$\sqrt{g}(\bar{y}_g - \bar{y}) = O_p(1), \quad (2.4.35)$$

so that by first fixing n and letting $g \rightarrow \infty$ (2.4.34) reduces to

$$\sqrt{n}(\bar{y}_g - \bar{Y}) = \sqrt{n}(\bar{y} - \bar{Y}) + o_p(1), \quad (2.4.36)$$

and the result of the theorem follows from condition C4 and Slutsky's theorem.

Proof of (2.4.35)

By Theorem 2.3, $E(\bar{y}_1^* | s_0) = \bar{y}$ is finite. Also $V(\bar{y}_g | s_0) = \sigma_*^2/g$, where $\sigma_*^2 = V(\bar{y}_1^* | s_0)$ is finite by C3. Now since $(\bar{y}_g - \bar{y})$ has the same order of magnitude as its conditional standard deviation, $\sigma_*/\sqrt{g}$, $(\bar{y}_g - \bar{y}) = O_p(g^{-\frac{1}{2}})$ by Chebyshev's inequality. Hence the proof of (2.4.35) follows. This completes the proof. ■

Theorem 2.8 *Under mild regularity conditions, (i) the variance of the inverse-sampling estimator $\bar{y}_g$, $V(\bar{y}_g)$, converges to the variance of the design-based estimator $\bar{y}$, $V(\bar{y})$,*

(ii) the inverse-sampling variance estimator, $\hat{V}(\bar{y}_g)$ is a consistent estimator of the design-based variance, $V(\bar{y})$, as $g \rightarrow \infty$ and $n \rightarrow \infty$.

Proof

(i) First note that $n\{V(\bar{y}_g) - V(\bar{y})\} = \frac{n}{g}EV(\bar{y}_1^*|s_0)$ by the standard relationship between conditional and unconditional variances and under the assumptions of Theorem 2.3.

The proof follows from the finiteness of $EV(\bar{y}_1^*|s_0)$ and letting $g \rightarrow \infty$ followed by $n \rightarrow \infty$.

(ii) Write

$$n\{\hat{V}(\bar{y}_g) - V(\bar{y})\} = \frac{n}{g}g\{\hat{V}(\bar{y}_g) - \hat{V}(\bar{y})\} + n\{\hat{V}(\bar{y}) - V(\bar{y})\}.$$

By Theorem 2.4, $g\{\hat{V}(\bar{y}_g) - \hat{V}(\bar{y})\} = o_p(1)$ as $g \rightarrow \infty$. Also by assumption C2, $n\{\hat{V}(\bar{y}) - V(\bar{y})\} = o_p(1)$. Hence, the result follows by letting $g \rightarrow \infty$ followed by $n \rightarrow \infty$.

■

2.4.3 Approximate Matching

In Section 2.2 we noted that exact matching with SRS is difficult to implement when the original sampling design involves clusters. We proposed several approximate matching methods to overcome this difficulty. In this subsection we study the properties of the approximate matching methods.

(i) Unistage Cluster Sampling

In Section 2.2.2, we considered the case of equal clusters, M , and proposed to select one element at random from each sample cluster i ($= 1, \dots, k$) selected with

equal probabilities and without replacement. To study some of the properties of the proposed approximate matching methods, we consider the estimation of a finite population total, Y . The estimator of Y is given by $\hat{Y} = (K/k) \sum_{i=1}^k Y_i$, where Y_i is the i -th sample cluster total and K is the number of population clusters. The variance of $\hat{Y}$ is given by

$$\hat{V}(\hat{Y}) = \frac{K^2}{k} \left(1 - \frac{k}{K}\right) \frac{1}{k-1} \sum_{i=1}^k \left(Y_i - \frac{1}{k} \sum_{i=1}^k Y_i\right)^2.$$

The inverse-sampling estimator of Y is given by $\hat{Y}_g = g^{-1} \sum_{j=1}^g \hat{Y}_j^*$, where $\hat{Y}_j^* = KM\bar{y}_j^*$ denoting the estimator of Y from the j -th inverse-sample. The inverse-sampling variance estimator of $\hat{Y}_g$ is given by

$$\hat{V}_g = \hat{V}(\hat{Y}_g) = \frac{1}{g} \sum_{j=1}^g \hat{V}(\hat{Y}_j^*) - \frac{1}{g} \sum_{j=1}^g (\hat{Y}_j^* - \hat{Y}_g)^2.$$

Theorem 2.9 *Under the approximate matching proposed for the unistage cluster sampling, (i) $\hat{Y}_\infty = \lim_{g \rightarrow \infty} \hat{Y}_g = \hat{Y}$ (ii) $\hat{V}_\infty = \lim_{g \rightarrow \infty} \hat{V}_g = (1 - k/K)^{-1} \hat{V}(\hat{Y})$.*

The statement of the theorem above means that, the approximate matching method preserves the original estimator $\hat{Y}$ in the limit whereas the variance estimator of $\hat{Y}$ is not preserved. This means that $\hat{V}_\infty$ leads to overestimation of the variance if the sampling fraction k/K is not small.

Proof of (i) and (ii) of Theorem 2.9

(i) We observe that $\hat{Y}_1^*, \dots, \hat{Y}_g^*$ are independent identically distributed bounded random variables. Therefore,

$$\hat{Y}_\infty = E(\hat{Y}_1^* | s_0) = (KM)E\left(\frac{1}{k} \sum_{i \in s_1^*} y_i^* | s_0\right) = KM \frac{1}{kM} \sum_{i=1}^k \sum_{j=1}^M Y_{ij} = \frac{K}{k} \sum_{i=1}^k Y_i = \hat{Y}.$$

(ii) We first define $\bar{y}_i = M^{-1} \sum_{j=1}^M y_{ij}$ and $\bar{\bar{y}} = k^{-1} \sum_{i=1}^k \bar{y}_i$. The inverse-sampling variance in the limit,

$$\begin{aligned}
 \hat{V}(\hat{Y}_\infty) &= E \{ \hat{V}(\hat{Y}_1^*) | s_0 \} - V(\hat{Y}_1^* | s_0) \\
 &= (KM)^2 [E \{ \hat{V}(\bar{y}_1^*) | S_0 \} - E(\bar{y}_1^* | s_0) + E^2(\bar{y}_1^* | s_0)] \\
 &= (KM)^2 \left[E \left\{ \frac{1}{k(k-1)} \left(\sum_{i=1}^k y_{1i}^{*2} - k\bar{y}_1^{*2} \right) | s_0 \right\} - E(\bar{y}_1^{*2} | s_0) + E^2(\bar{y}_1^* | s_0) \right] \\
 &= (KM)^2 \left[\frac{1}{k(k-1)} \left\{ \sum_{i=1}^k \bar{y}_i^2 - \left(\sum_{i=1}^k \bar{y}_i \right)^2 \right\} + \bar{\bar{y}}^2 \right] \\
 &= \frac{(KM)^2}{k(k-1)} \sum_{i=1}^k (\bar{y}_i - \bar{\bar{y}})^2 \\
 &= \left(1 - \frac{k}{K} \right)^{-1} \hat{V}(\hat{Y})
 \end{aligned}$$

■

(ii) Two-Stage Cluster Sampling

Consider the case of two-stage cluster sampling with equal cluster sizes, M , and SRS without replacement in both stages. As noted in Section 2.2.3, exact matching is difficult to implement. The approximate inverse sampling method consists of selecting one element at random from the m sample elements in each sample cluster i ($= 1, \dots, k$). Denote the values of the elements by $y'_1, \dots, y'_k$. The H-T estimator of Y is given by $\hat{Y} = (K/k) \sum_{i=1}^k \hat{Y}_i$, where $\hat{Y}_i = M\bar{y}_i$ and $\bar{y}_i$ is the sample mean of the i -th sample cluster. The repeated inverse-sampling estimator, based on approximate matching, is given by $\hat{Y}_g = g^{-1} \sum_{j=1}^g \hat{Y}_j^*$, where $\hat{Y}_1^* = (N/k) \sum_{i=1}^k y'_i$. It is easy to verify that $\hat{Y}_\infty = \hat{Y}$ so that approximate matching preserves the original estimator $\hat{Y}$

in the limit. The variance estimator of $\hat{Y}$ is given by

$$\hat{V} = N^2 \left\{ \frac{1}{k} \left(1 - \frac{k}{K} \right) s_{1y}^2 + \frac{k}{K} \left(1 - \frac{m}{M} \right) \frac{1}{km} s_{2y}^2 \right\}, \quad (2.4.37)$$

where $s_{1y}^2 = \sum_i (\bar{y}_i - \bar{y})^2 / (k - 1)$, $s_{2y}^2 = \sum_i s_{2i}^2 / k$ with s_{2i}^2 denoting the sample variance in the i -th cluster, $\bar{y}_i$ is the i -th cluster sample mean and $\bar{y} = \sum_i \bar{y}_i / k$ is the overall sample mean (see Cochran, 1977, pages 276-278). Following similar arguments as in unistage cluster sampling, the inverse-sampling variance estimator $\hat{V}_g$ is easily shown to go in the limit to

$$\hat{V}_\infty = N^2 \frac{1}{k} s_{1y}^2 \quad (2.4.38)$$

as $g \rightarrow \infty$. It follows from (2.4.37) and (2.4.38) that

$$\begin{aligned} \frac{\hat{V}}{\hat{V}_\infty} &= 1 - \frac{k}{K} \left\{ 1 - \left(1 - \frac{m}{M} \right) \frac{1}{km} \frac{s_{2y}^2}{s_{1y}^2} \right\} \\ &\approx 1 - \frac{k}{K}, \end{aligned} \quad (2.4.39)$$

because the neglected term in (2.4.39) is of order $(mK)^{-1}$. It follows that $\hat{V}_\infty$ again leads to overestimation of the variance if the sampling fraction k/K is not small.

2.5 Improving the Efficiency of $\hat{Y}_g$

Suppose the full-sample estimator $\hat{X}$ of the population total of an auxiliary variable x is available. Then inverse sampling may be regarded as a special case of a two-phase sampling: phase one corresponds to the original full-sample s_0 giving $\hat{X}$, and phase two consists of g subsamples $\{(y_i, x_i); i \in s_j^*\}$, $j = 1, \dots, g$. The phase one information, $\hat{X}$, may be used to improve the efficiency of the inverse-sampling estimator $\hat{Y}_g$

of the population total Y by the use of a two-phase ratio estimator of a population total.

An inverse-sampling ratio estimator, $\hat{Y}_{rg}$, of Y is given by

$$\hat{Y}_{rg} = \frac{\hat{Y}_g}{\hat{X}_g} \hat{X} = \hat{R}_g \hat{X}, \quad (2.5.40)$$

where $\hat{Y}_g = \{N/(gm)\} \sum_{j=1}^g \sum_{i \in s_j^*} y_i = g^{-1} \sum_{j=1}^g \hat{Y}_j^*$, $\hat{X}_g = \{N/(gm)\} \sum_{j=1}^g \sum_{i \in s_j^*} x_i = g^{-1} \sum_{j=1}^g \hat{X}_j^*$, and $\hat{R} = \hat{Y}_g / \hat{X}_g$.

Using the two-phase approach, a linearization approximation to the variance of $\hat{Y}_{rg}$ is given by

$$V(\hat{Y}_{rg}) \cong V(\hat{Y}_g) - \frac{1}{g} \left\{ EV(\hat{Y}_1^* | s_0) - EV(E_1^* | s_0) \right\}, \quad (2.5.41)$$

where $E_1^* = \hat{Y}_1^* - R\hat{X}_1^*$, $R = Y/X$ is a ratio based on all population elements. The result in (2.5.41) shows that the inverse-sampling ratio estimator $\hat{Y}_{rg}$ is asymptotically more efficient than the inverse-sampling estimator, $\hat{Y}_g$, of the population total Y , if the study variable, y is proportionally related to the auxiliary variable, x . The result (2.5.41) suggests that one should report the full-sample estimator $\hat{X}$ with the microdata set consisting of g subsamples $\{(y_i, x_i), i \in s_j^*\} j = 1, \dots, g$.

From (2.5.41) we easily obtain an inverse-sampling variance estimator of $\hat{Y}_{rg}$ as

$$\hat{V}(\hat{Y}_{rg}) = \hat{V}(\hat{Y}_g) - \frac{1}{g} \left\{ \frac{1}{g-1} \sum_{j=1}^g (\hat{Y}_j^* - \hat{Y}_g)^2 - \frac{1}{g-1} \sum_{j=1}^g (\hat{E}_j^* - \hat{E}_g)^2 \right\}, \quad (2.5.42)$$

where $\hat{E}_j^* = \hat{Y}_j^* - \hat{R}_g \hat{X}_j^*$, $\hat{E}_g = g^{-1} \sum_{j=1}^g \hat{E}_j^*$, and $\hat{V}(\hat{Y}_g)$ is the variance estimator of $\hat{Y}_g$ given in (2.3.14).

2.6 Inverse-Sampling Generalized Regression Estimator of a Total

The generalized regression estimator is another estimator that attempts to use auxiliary information to improve the precision of an estimator of population total. Suppose $\mathbf{x}_k = (x_{k1}, \dots, x_{kp})'$ is the $p \times 1$ vector of auxiliary variables for the k -th population element, and y_k is the k -th element of the corresponding study variable, y . Assume further that the total X of $\mathbf{x}_k$'s is known.

Even though the focus is the generalized regression estimator, we first study the properties of the closely related difference estimator and then turn our attention to the generalized regression estimator.

2.6.1 Inverse-Sampling Difference Estimator

Suppose a sample s_0 is drawn from the population U by a sampling design $p_0(s_0)$.

Then the full-sample difference estimator is given by

$$\hat{Y}_d = \sum_{k \in U} y_k^0 + \sum_{k \in s_0} \frac{d_k}{\pi_k} = \mathbf{X}' \mathbf{B}^0 + \sum_{k \in s_0} \frac{d_k}{\pi_k}, \quad (2.6.43)$$

where $y_k^0 = \mathbf{x}_k' \mathbf{B}^0$, $\mathbf{B}^0 = (B_1^0, \dots, B_p^0)'$ is specified, and $d_k = y_k - y_k^0$. A full-sample variance estimator of $\hat{Y}_d$ is given by

$$\hat{V}(\hat{Y}_d) = \sum_{i,j \in s_0} \sum \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{d_i}{\pi_i} \frac{d_j}{\pi_j}. \quad (2.6.44)$$

The inverse-sampling difference estimator we propose is given by

$$\hat{Y}_{dg} = \mathbf{X}' \mathbf{B}^0 + \frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j^*} \frac{d_k}{\pi_k^*}. \quad (2.6.45)$$

Note that the data file consisting of g inverse samples $\{(y_i, x_i), i \in s_j^*, j, \dots, g\}$ should include the totals $\mathbf{X}$ in order to implement $\hat{Y}_{dg}$. The inverse-sampling variance estimator of $\hat{Y}_{dg}$ is given by

$$\hat{V}(\hat{Y}_{dg}) = \frac{1}{g} \sum_{j=1}^g \hat{V}_{dj}^* - \frac{1}{g} \sum_{j=1}^g (\hat{D}_j^* - \hat{D}_g)^2, \quad (2.6.46)$$

where

$$\hat{V}_{dj}^* = \sum_{k,l \in s_0} \frac{\pi_{kl}^* - \pi_k^* \pi_l^*}{\pi_{kl}^*} \frac{d_k d_l}{\pi_k^* \pi_l^*},$$

$$\hat{D}_j^* = \sum_{k \in s_j^*} d_k / \pi_k^* \text{ and } \hat{D}_g = g^{-1} \sum_{j=1}^g \hat{D}_j^*.$$

Theorem 2.10 establishes conditions under which the inverse-sampling difference estimator, $\hat{Y}_{dg}$ converges to the full-sample estimator $\hat{Y}_d$, as $g \rightarrow \infty$.

Theorem 2.10 *Let $\tilde{\pi}_i(s_0)$ be the conditional probability that the i -th unit is selected in the inverse-sample for a given initial sample, s_0 . Then the limiting inverse-sampling difference estimator $\hat{Y}_{d,\infty}$ is the full-sample difference estimator $\hat{Y}_d$, if and only if the conditional inclusion probabilities $\tilde{\pi}_i(s_0)$ are constant for all s_0 containing the i -th unit.*

Proof

The proof is similar to the proof of Theorem 2.3. ■

Theorem 2.11 *If $\hat{V}_{dj}^*$ is the Horvitz-Thompson (H-T) variance estimator of $\hat{Y}_{dj}^* = \mathbf{X}' \mathbf{B}^0 + \sum_{k \in s_j^*} d_k / \pi_k^*$ for the j -th subsample, then conditional on s_0 , $\hat{V}(\hat{Y}_{dg})$, converges to the Horvitz-Thompson (H-T) variance estimator of $\hat{Y}_d$ for the original design,*

as $g \rightarrow \infty$, if the conditional joint inclusion probabilities are constant for all s_0 containing a given pair (i, l) of units.

Proof

The proof is similar to the proof of Theorem 2.4.

■

2.6.2 Inverse-Sampling Regression Estimator

Suppose $(y_k, \mathbf{x}_k), k \in s_0$ is observed and $\mathbf{x}_k$ is also known for $k \in U - s_0$. Consider the “working” regression model ξ with the following assumptions (i) $y_1, \dots, y_N$ are realized values of independent random variables $Y_1, \dots, Y_N$; (ii) $E_\xi(Y_k) = \mathbf{x}_k' \boldsymbol{\beta}$ ($k = 1, \dots, N$) and (iii) $V_\xi(Y_k) = \sigma_k^2$ ($k = 1, \dots, N$), where E_ξ and V_ξ are the expected value and variance with respect to the model ξ , $\boldsymbol{\beta}$ and $\sigma_1, \dots, \sigma_N$ are model parameters.

The generalized least squares or weighted least square estimator of $\boldsymbol{\beta}$ is given by $\mathbf{B} = \hat{\boldsymbol{\beta}} = (\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k' / \sigma_k^2)^{-1} (\sum_{k \in U} \mathbf{x}_k y_k / \sigma_k^2) = \mathbf{T}^{-1} \mathbf{t}$ and its estimator from the full sample is $\widehat{\mathbf{B}} = (\sum_{k \in s_0} \mathbf{x}_k \mathbf{x}_k' / (\sigma_k^2 \pi_k))^{-1} (\sum_{k \in s_0} \mathbf{x}_k y_k / (\sigma_k^2 \pi_k)) = \widehat{\mathbf{T}}^{-1} \widehat{\mathbf{t}}$. Substituting $\widehat{\mathbf{B}}$ for $\mathbf{B}^0$ in the difference estimator we obtain the generalized regression (GREG) estimator of Y

$$\hat{Y}_{reg} = \sum_{k \in U} \hat{y}_k + \sum_{k \in s_0} \frac{e_{ks_0}}{\pi_k} = \sum_{k \in s_0} \frac{g_{ks_0} y_k}{\pi_k}, \quad (2.6.47)$$

where $\hat{y}_k = \mathbf{x}_k' \widehat{\mathbf{B}}$, $e_{ks_0} = y_k - \hat{y}_k$ and $g_{ks_0} = 1 + (\mathbf{X} - \widehat{\mathbf{X}})' \widehat{\mathbf{T}}^{-1} \mathbf{x}_k / \sigma_k^2$ is called the g-weight (Särndal et al., 1992, page 231).

A linearization full-sample variance estimator of $\hat{Y}_r$ is given by

$$\hat{V}(\hat{Y}_{reg}) = \sum_{k,l \in s_0} \sum \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{(g_{ks_0} e_{ks_0})}{\pi_k} \frac{(g_{ls_0} e_{ls_0})}{\pi_l}, \quad (2.6.48)$$

see Särndal et al. (1992, page 235).

An alternative variance estimator is given by

$$\hat{V}(\hat{Y}_{reg}) = \sum_{k,l \in s_0} \sum \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{e_{ks_0}}{\pi_k} \frac{e_{ls_0}}{\pi_l}. \quad (2.6.49)$$

The inverse-sampling generalized regression estimator we propose is given by

$$\hat{Y}_{reg,g} = \mathbf{X}' \hat{\mathbf{B}}_g + \frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j^*} \frac{e_k^*}{\pi_k^*} = \frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j^*} \frac{g_k^* y_k}{\pi_k^*}, \quad (2.6.50)$$

where $e_k^* = y_k^* - \mathbf{x}_k^{*'} \hat{\mathbf{B}}_g$, $g_k^* = 1 + (\mathbf{X} - \hat{\mathbf{X}}_g)' \hat{\mathbf{T}}^{*-1} \mathbf{x}_k / \sigma_k^2$ and

$$\hat{\mathbf{B}}_g = \left(\frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j^*} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2} \right)^{-1} \left(\frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j^*} \frac{\mathbf{x}_k y_k}{\sigma_k^2} \right) = \hat{\mathbf{T}}^{*-1} \hat{\mathbf{t}}^*.$$

The inverse-sampling variance estimator of $\hat{Y}_{reg,g}$ is given by

$$\hat{V}(\hat{Y}_{reg,g}) = \frac{1}{g} \sum_{j=1}^g \hat{V}_{reg,j}^* - \frac{1}{g} \sum_{j=1}^g \left(\hat{U}_j^*(\hat{\mathbf{B}}_g) - \bar{U}(\hat{\mathbf{B}}_g) \right)^2, \quad (2.6.51)$$

where

$$\hat{V}_{reg,j}^* = \sum_{k,l \in s_j^*} \sum \frac{\pi_{kl}^* - \pi_k^* \pi_l^*}{\pi_{kl}^*} \frac{e_k^*}{\pi_k^*} \frac{e_l^*}{\pi_l^*},$$

where $e_k^* = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_g$, $\hat{U}_j^*(\hat{\mathbf{B}}_g) = \sum_{k \in s_j^*} e_k^* / \pi_k^*$ and $\bar{U}(\hat{\mathbf{B}}_g) = 1/g \sum_{j=1}^g \hat{U}_j^*(\hat{\mathbf{B}}_g)$.

Theorem 2.12 provides conditions under which the inverse-sampling generalized regression estimator, $\hat{Y}_{reg,g}$, converges to the full-sample generalized regression estimator, $\hat{Y}_{reg}$.

Theorem 2.12 *Let $\tilde{\pi}_i(s_0)$ be the conditional probability that the i -th unit is selected in the inverse-sample for a given initial sample s_0 . Then the inverse-sampling generalized regression estimator, $\hat{Y}_{reg,g}$, converges almost surely to the full-sample generalized regression estimator, if and only if the conditional inclusion probabilities are constant for all s_0 containing the i -th unit.*

Proof

We note that

$$\begin{aligned}\hat{Y}_{reg,g} &= \mathbf{X}'\hat{\mathbf{B}}_g + \frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j^*} \frac{e_k^*}{\pi_k^*} \\ &= \mathbf{X}'\hat{\mathbf{B}}_g + \frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j^*} \frac{e_{ks_0}}{\pi_k^*} + \hat{\mathbf{X}}_g'(\hat{\mathbf{B}} - \hat{\mathbf{B}}_g).\end{aligned}$$

The proof of the theorem follows by noting that conditional on s_0 , as $g \rightarrow \infty$, (i) $\hat{\mathbf{B}}_g$ converges almost surely to $\hat{\mathbf{B}}$; (ii) the second term converges almost surely to $\sum_{k \in s_0} e_{ks_0}/\pi_k$ and (iii) the third term converges almost surely to 0, noting that $\mathbf{X}_g$ and $\hat{\mathbf{B}}_g$ converge almost surely to $\hat{\mathbf{X}}$ and $\hat{\mathbf{B}}$ respectively, as $g \rightarrow \infty$. ■

Theorem 2.13 *Suppose $\hat{V}_j^*$ is the Horvitz-Thompson (H-T) variance estimator of $\hat{U}_j^*(\hat{\mathbf{B}}_g)$ for the j -th subsample, then conditional on s_0 , $\hat{V}(\hat{Y}_{reg,g})$ converges in probability to the Horvitz-Thompson (H-T) variance estimator of $\hat{Y}_{reg}$ (given in 2.6.49), for the original design, as $g \rightarrow \infty$ provided the conditional joint inclusion probabilities are constant for all s_0 containing a given pair (i, l) of units.*

Proof

Using the fact that $\widehat{\mathbf{B}}_g \rightarrow \widehat{\mathbf{B}}$ as $g \rightarrow \infty$, we can write

$$\widehat{V}(\widehat{Y}_{reg,g}) \approx \frac{1}{g} \sum_{j=1}^g \widehat{V}_j^*(\widehat{\mathbf{B}}) - \frac{1}{g} \sum_{j=1}^g \{\widehat{U}_j^*(\widehat{\mathbf{B}}) - \overline{U}(\widehat{\mathbf{B}})\}^2.$$

Now conditional on s_0 , as $g \rightarrow \infty$, the first two terms above converge almost surely to $E\{\widehat{V}_1^*(\widehat{\mathbf{B}})|s_0\} - V\{\widehat{U}_1^*(\widehat{\mathbf{B}})|s_0\}$. The proof of the theorem follows by showing that $E\{\widehat{V}_1^*(\widehat{\mathbf{B}})|s_0\} - V\{\widehat{U}_1^*(\widehat{\mathbf{B}})|s_0\}$ is equal to (2.6.49). The proof of this latter assertion is similar to the proof of Theorem 2.4. ■

If we substitute $(g_k^* e_k^*)$ for e_k^* in the inverse-sampling variance estimator (2.6.51), this alternative inverse-sampling variance estimator can be shown to converge to the full-sample variance estimator (2.6.48) following similar arguments in the proof of Theorem 2.13.

2.7 Estimating Quadratic Population Parameters

In this section we study the estimation of quadratic census population parameters which are linear functions of $y_i y_j$ or $(y_i - y_j)^2$ or $(x_i - x_j)(y_i - y_j)$. The quadratic census population parameters are given by

$$T_2 = \sum_{i=1}^N \sum_{j=i}^N c_{ij} y_i y_j, \quad (2.7.52)$$

$$T_3 = \sum_{i=1}^{N-1} \sum_{j=i+1}^N b_{ij} (y_i - y_j)^2, \quad (2.7.53)$$

and

$$T_4 = \sum_{i < j}^N \sum_{i < j}^N b_{ij}(x_i - x_j)(y_i - y_j) \quad (2.7.54)$$

respectively. A population variance, $S_y^2 = (N-1)^{-1} \sum_{i=1}^N (y_i - \bar{Y})^2$ is a special case of T_3 and a population covariance between x and y , $S_{xy} = (N-1)^{-1} \sum_{i=1}^N (x_i - \bar{X})(y_i - \bar{Y})$ is an example of T_4 .

The Horvitz-Thompson unbiased estimators of T_2 , T_3 and T_4 based on the full-sample are given by

$$\hat{T}_2 = \sum_{i,j \in s_0; i \leq j} \frac{c_{ij} y_i y_j}{\pi_{ij}}, \quad (2.7.55)$$

$$\hat{T}_3 = \sum_{i,j \in s_0; i < j} \frac{b_{ij}(y_i - y_j)^2}{\pi_{ij}}, \quad (2.7.56)$$

and

$$\hat{T}_4 = \sum_{i,j \in s_0; i < j} \frac{b_{ij}(x_i - x_j)(y_i - y_j)}{\pi_{ij}} \quad (2.7.57)$$

respectively. The full-sample variance estimators of $\hat{T}_2$, $\hat{T}_3$ and $\hat{T}_4$ are given by

$$\hat{V}(\hat{T}_2) = \sum_{i \leq j} \sum_{l \leq m \in s_0} \sum_{i \leq j} \sum_{l \leq m \in s_0} c_{ij} c_{lm} \frac{\pi_{ijlm} - \pi_{ij} \pi_{lm}}{\pi_{ijlm} \pi_{ij} \pi_{lm}} y_i y_j y_l y_m, \quad (2.7.58)$$

$$\hat{V}(\hat{T}_3) = \sum_{i < j} \sum_{l < m \in s_0} \sum_{i < j} \sum_{l < m \in s_0} b_{ij} b_{lm} \frac{\pi_{ijlm} - \pi_{ij} \pi_{lm}}{\pi_{ijlm} \pi_{ij} \pi_{lm}} (y_i - y_j)^2 (y_l - y_m)^2, \quad (2.7.59)$$

and

$$\hat{V}(\hat{T}_4) = \sum_{i < j} \sum_{l < m \in s_0} \sum_{i < j} \sum_{l < m \in s_0} \frac{b_{ij} b_{lm} (\pi_{ijlm} - \pi_{ij} \pi_{lm})}{\pi_{ijlm} \pi_{ij} \pi_{lm}} (x_i - x_j)(y_i - y_j)(x_l - x_m)(y_l - y_m), \quad (2.7.60)$$

where π_{ijlm} is the fourth-order inclusion probability in s_0 .

The inverse-sampling estimators of T_2 , T_3 and T_4 are given by

$$\hat{T}_{2g} = \frac{1}{g} \sum_{j=1}^g \hat{T}_{2j}^* = \frac{1}{g} \sum_{j=1}^g \sum_{i,l \in s_j^*; i \leq l} \frac{c_{il} y_i y_l}{\pi_{il}^*}, \quad (2.7.61)$$

$$\hat{T}_{3g} = \frac{1}{g} \sum_{j=1}^g \hat{T}_{3j}^* = \frac{1}{g} \sum_{j=1}^g \sum_{i,l \in s_j^*; i < l} \frac{b_{il} (y_i - y_l)^2}{\pi_{il}^*}, \quad (2.7.62)$$

and

$$\hat{T}_{4g} = \frac{1}{g} \sum_{j=1}^g \hat{T}_{4j}^* = \frac{1}{g} \sum_{j=1}^g \sum_{i,l \in s_j^*; i \leq l} \frac{b_{il} (x_i - x_l)(y_i - y_l)}{\pi_{il}^*}, \quad (2.7.63)$$

respectively. Theorem 2.14 below establishes conditions under which the inverse-sampling estimators, $\hat{T}_{2g}$, $\hat{T}_{3g}$ and $\hat{T}_{4g}$ converge to the H-T estimators, $\hat{T}_2$, $\hat{T}_3$ and $\hat{T}_4$ for the original full-sample as $g \rightarrow \infty$.

Theorem 2.14 *Let $\bar{\pi}_{il}(s_0)$ be the conditional joint probability that the pair (i, l) of units is selected in the subsample for a given initial sample, s_0 . Suppose that $\hat{\theta}_j^* = \hat{T}_{kj}^*$ ($k = 2, 3, 4$) is the H-T estimator of a finite population quadratic function $\theta = T_k$ ($k = 2, 3, 4$) for the j -th subsample. Then the limiting inverse-sampling estimator, $\hat{\theta}_\infty^* = \hat{T}_{k,\infty}$, will be the H-T estimator, $\hat{T}_k$, for the original design if and only if the conditional joint inclusion probabilities $\bar{\pi}_{il}(s_0)$ are constant for all s_0 containing a given pair (i, l) of units, i.e., $\bar{\pi}_{il}(s_0) = \bar{\pi}_{il}$ for all $s_0 \supset \{i, l\}$.*

Proof

We would show that $\hat{T}_{2,\infty} = \hat{T}_2$, $\hat{T}_{3,\infty} = \hat{T}_3$ and $\hat{T}_{4,\infty} = \hat{T}_4$.

Case 1. $\hat{T}_{2,\infty} = \hat{T}_2$

For

$$\hat{T}_{2j}^* = \sum_{i,l \in s_j^*; i \leq l} \frac{c_{il} y_i y_l}{\pi_{il}^*} = \sum_{i,l \in s_0; i \leq l} \frac{c_{il} y_i y_l}{\pi_{il}^*} I_{(i,l),j}^*(s_0),$$

where $I_{(i,l),j}^*(s_0)$ takes the value 1 if the pair (i,l) is included in the j -th subsample s_j^* and 0 otherwise, and π_{il}^* is the corresponding (unconditional) joint selection probability. Thus

$$\hat{T}_{2,\infty} = E(\hat{T}_{21}^* | s_0) = \sum_{i,l \in s_0; i \leq l} \frac{c_{il} y_i y_l}{\pi_{il}^*} \tilde{\pi}_{il}(s_0).$$

This is equal to

$$\hat{T}_2 = \sum_{i,l \in s_0; i \leq l} \frac{c_{il} y_i y_l}{\pi_{il}},$$

the H-T estimator for the original design, if and only if $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il} = \pi_{il}^* / \pi_{il}$.

Case 2. $\hat{T}_{3,\infty} = \hat{T}_3$

For

$$\hat{T}_{3j}^* = \sum_{i,l \in s_j^*; i < l} \frac{b_{il}(y_i - y_l)^2}{\pi_{il}^*} = \sum_{i,l \in s_0; i < l} \frac{b_{il}(y_i - y_l)^2}{\pi_{il}^*} I_{(i,l),j}^*(s_0),$$

where $I_{(i,l),j}^*(s_0)$ is as defined in Case 1 above. Hence

$$\hat{T}_{3,\infty} = E(\hat{T}_{31}^* | s_0) = \sum_{i,l \in s_0; i < l} \frac{b_{il}(y_i - y_l)^2}{\pi_{il}^*} \tilde{\pi}_{il}(s_0).$$

This is equal to $\hat{T}_3$ provided $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il} = \pi_{il}^* / \pi_{il}$.

Case 3. $\hat{T}_{4,\infty} = \hat{T}_4$

For

$$\hat{T}_{4j}^* = \sum_{i,l \in s_j^*; i < l} \frac{b_{il}(x_i - x_l)(y_i - y_l)}{\pi_{il}^*} = \sum_{i,l \in s_0; i < l} \frac{b_{il}(x_i - x_l)(y_i - y_l)}{\pi_{il}^*} I_{(i,l),j}^*(s_0),$$

where $I_{(i,l),j}^*(s_0)$ is as defined in Case 1 above. Hence

$$\hat{T}_{4,\infty} = E(\hat{T}_{41}^* | s_0) = \sum_{i,l \in s_0; i < l} \sum \frac{b_{il}(x_i - x_l)(y_i - y_l)}{\pi_{il}^*} \tilde{\pi}_{il}(s_0).$$

This is equal to $\hat{T}_4$ provided $\tilde{\pi}_{il}(s_0) = \tilde{\pi}_{il} = \pi_{il}^* / \pi_{il}$.

■

The inverse-sampling variance estimator of $\hat{T}_{kg}$, $k = 2, 3, 4$ is given by

$$\hat{V}(\hat{T}_{kg}) = \frac{1}{g} \sum_{j=1}^g \hat{V}_{kj}^* - \frac{1}{g} \sum_{j=1}^g (\hat{T}_{kj}^* - \hat{T}_{kg})^2, \quad k = 2, 3, 4, \quad (2.7.64)$$

where for $j = 1, \dots, g$,

$$\hat{V}_{2j}^* = \sum_{i \leq k} \sum_{l \leq m \in s_j^*} \sum c_{ik} c_{lm} \frac{\pi_{iklm}^* - \pi_{ik}^* \pi_{lm}^*}{\pi_{iklm}^* \pi_{ik}^* \pi_{lm}^*} y_i y_k y_l y_m, \quad (2.7.65)$$

$$\hat{V}_{3j}^* = \sum_{i < k} \sum_{l < m \in s_j^*} \sum b_{ik} b_{lm} \frac{\pi_{iklm}^* - \pi_{ik}^* \pi_{lm}^*}{\pi_{iklm}^* \pi_{ik}^* \pi_{lm}^*} (y_i - y_k)^2 (y_l - y_m)^2, \quad (2.7.66)$$

and

$$\hat{V}_{4j}^* = \sum_{i < k} \sum_{l < m \in s_j^*} \sum \frac{b_{ik} b_{lm} (\pi_{iklm}^* - \pi_{ik}^* \pi_{lm}^*)}{\pi_{iklm}^* \pi_{ik}^* \pi_{lm}^*} (x_i - x_k)(y_i - y_k)(x_l - x_m)(y_l - y_m), \quad (2.7.67)$$

where π_{iklm}^* is the fourth-order inclusion probability in s^* .

The following theorem establishes conditions under which the inverse-sampling variance estimator, $\hat{V}(\hat{T}_{kg})$, $k = 2, 3, 4$ converge to the full-sample variance estimator $\hat{V}(\hat{T}_k)$.

Theorem 2.15 *If $\hat{V}_{kj}^*$ is the Horvitz-Thompson (H-T) variance estimator of $\hat{T}_{kj}^*$ for the j -th subsample ($k = 2, 3, 4$), then conditional on s_0 , $\hat{V}(\hat{T}_{kg})$ converges to the*

full-sample Horvitz-Thompson (H-T) variance estimator of $\hat{T}_k$ as $g \rightarrow \infty$, if the conditional joint inclusion probabilities are constant for all s_0 containing (i, k, l, m) of units, i.e., $\tilde{\pi}_{iklm}(s_0)$ for all $s_0 \supset \{i, k, l, m\}$.

Proof

We give the proof for $k = 3$ which easily extends to $k = 2$ and $k = 4$. We first note that conditional on s_0 , $\hat{V}(\hat{T}_{3g})$ converges almost surely to $E\{\hat{V}(\hat{T}_{31}^*)|s_0\} - V(\hat{T}_{31}^*|s_0)$, as $g \rightarrow \infty$. It is therefore sufficient to show that $E\{\hat{V}(\hat{T}_{31}^*)|s_0\} - V(\hat{T}_{31}^*|s_0) = \hat{V}(\hat{T}_3)$.

Now

$$\begin{aligned} & E\{\hat{V}(\hat{T}_{31}^*)|s_0\} \\ &= \sum_{i < k} \sum_{l < m \in s_0} \sum b_{ik} b_{lm} \frac{\pi_{iklm}^* - \pi_{ik}^* \pi_{lm}^*}{\pi_{iklm}^* \pi_{ik}^* \pi_{lm}^*} \tilde{\pi}_{iklm} (y_i - y_k)^2 (y_l - y_m)^2, \\ &= \sum_{i < k} \sum_{l < m \in s_0} \sum b_{ik} b_{lm} \left(\frac{\tilde{\pi}_{iklm}}{\pi_{ik}^* \pi_{lm}^*} - \frac{1}{\pi_{iklm}} \right) (y_i - y_k)^2 (y_l - y_m)^2 \quad (2.7.68) \end{aligned}$$

and

$$\begin{aligned} & V(\hat{T}_{31}^*|s_0) \\ &= \sum_{i < k} \sum_{l < m \in s_0} \sum b_{ik} b_{lm} \frac{\tilde{\pi}_{iklm} - \tilde{\pi}_{ik} \tilde{\pi}_{lm}}{\pi_{ik}^* \pi_{lm}^*} (y_i - y_k)^2 (y_l - y_m)^2 \\ &= \sum_{i < k} \sum_{l < m \in s_0} \sum b_{ik} b_{lm} \left(\frac{\tilde{\pi}_{iklm}}{\pi_{ik}^* \pi_{lm}^*} - \frac{1}{\pi_{iklm}} \right) (y_i - y_k)^2 (y_l - y_m)^2. \quad (2.7.69) \end{aligned}$$

From (2.7.68) and (2.7.69) we obtain

$$\sum_{i < k} \sum_{l < m \in s_0} \sum b_{ik} b_{lm} \frac{\pi_{iklm} - \pi_{ik} \pi_{lm}}{\pi_{iklm} \pi_{ik} \pi_{lm}} (y_i - y_k)^2 (y_l - y_m)^2 = \hat{V}(\hat{T}_3).$$

Hence the proof. ■

Chapter 3

Estimation of Complex Parameters

3.1 Estimating Equations Approach

In this chapter, we study a combined estimating equation approach to inverse sampling. This approach permits asymptotically valid inferences on nonlinear parameters such as ratios and census linear regression and logistic regression parameters. As noted in Section 2.3, the separate inverse-sampling estimator $\hat{\theta}_g = \sum_{j=1}^g \hat{\theta}_j^* / g$ has exactly the same bias as $\hat{\theta}_1^*$, and the bias of $\hat{\theta}_1^*$ is of order m^{-1} , where m is the subsample size. As a result, the bias of $\hat{\theta}_g$ can be appreciable because m is usually very much smaller than the original sample size n . In fact, m could be as small as 2 for stratified two-stage cluster sampling designs with two sample clusters in each stratum. Moreover, for logistic regression parameters and other complex parameters, one requires separate iterative solutions of g estimating equations to obtain the subsample estimators, $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$. As a result, the implementation of $\hat{\theta}_g$, and the associated inverse-sampling variance estimator $\hat{V}_g$, given by (2.3.13), could become computation-

ally very cumbersome when the number of subsamples, g , is large. We avoid some of these difficulties by using a combined estimating equation approach. Section 3.2 provides basic properties of the combined estimating equation (CEE) method. We extend this method to study post-stratified estimation in Section 3.3, analysis of categorical survey data in Section 3.4, and quasi-score tests with survey data in Section 3.5.

3.2 Basic Properties of the CEE Method

3.2.1 Full-sample estimating equations

Consider the finite population parameter vector θ_N defined as the solution to the census estimating equations:

$$U(\theta) = \sum_{k \in U} \mathbf{u}_k(\theta) = \mathbf{0}, \quad (3.2.1)$$

where $\sum_{k \in U}$ denotes the summation over the finite population U of size N , and the estimating functions $\mathbf{u}_k(\theta)$ are suitably chosen (Binder 1983; Godambe and Thompson 1986). For example, consider the scalar case of (3.2.1) and let $u_k(\theta) = y_k - \theta$ in (3.2.1). This gives the population mean $\theta_N = \bar{Y}$. Similarly, letting $u_k(\theta) = y_k - \theta x_k$ we get the ratio of totals: $\theta_N = R = Y/X$. The choice $\mathbf{u}_k(\theta) = \mathbf{x}_k(y_k - \mu_k(\theta))$ with $\mu_k(\theta) = \mathbf{x}_k' \theta$ gives the least squares regression parameter

$$\theta_N = \left(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in U} \mathbf{x}_k y_k.$$

The choice $\mathbf{u}_k(\theta) = \mathbf{x}_k(y_k - \mu_k(\theta))$ with $\log[\mu_k(\theta)/(1 - \mu_k(\theta))] = \mathbf{x}_k' \theta$ gives the logistic regression parameter θ_N . Kovacevic and Binder (1997) give estimating functions,

$u_k(\theta)$, that lead to various measures of income inequality, such as the Gini index and the polarization index.

The full-sample estimating equations are given by

$$\widehat{U}(\theta) = \sum_{k \in s_0} w_k u_k(\theta) = 0, \quad (3.2.2)$$

where w_k is the survey weight attached to $k \in s_0$; in particular, $w_k = 1/\pi_k$ if the H-T estimator of $U(\theta)$ is used. The solution to (3.2.2) gives the survey-based estimator $\hat{\theta}$ which, in general, is nonlinear and hence biased. We assume that the size of the original sample, s_0 , is large enough to neglect the bias of $\hat{\theta}$. For logistic regression and other complex cases, it is necessary to solve (3.2.2) iteratively to obtain the solution $\hat{\theta}$. The Newton-Raphson (N-R) algorithm is commonly used to solve (3.2.2). The r -th step of the N-R algorithm is given by

$$\hat{\theta}^{(r)} = \hat{\theta}^{(r-1)} + \left\{ - \frac{\partial \widehat{U}(\theta)}{\partial \theta'} \bigg|_{\theta = \hat{\theta}^{(r-1)}} \right\}^{-1} \widehat{U}(\hat{\theta}^{(r-1)}), \quad (3.2.3)$$

where $\hat{\theta}^{(r-1)}$ is the value of θ obtained at the $(r-1)$ -th iteration, and $\widehat{U}(\hat{\theta}^{(r-1)})$ is the value of $\widehat{U}(\theta)$ evaluated at $\theta = \hat{\theta}^{(r-1)}$. Iterating the N-R algorithm to convergence produces the estimator $\hat{\theta}$.

Under regularity conditions, Binder (1983) obtained a Taylor linearization variance estimator of $\hat{\theta}$ as

$$\widehat{V}_L(\hat{\theta}) = \left\{ \frac{\partial \widehat{U}(\theta)}{\partial \theta'} \bigg|_{\theta = \hat{\theta}} \right\}^{-1} \widehat{V} \{ \widehat{U}(\hat{\theta}) \} \left\{ \frac{\partial \widehat{U}(\theta)}{\partial \theta'} \bigg|_{\theta = \hat{\theta}} \right\}^{-1}, \quad (3.2.4)$$

where $\widehat{V} \{ \widehat{U}(\hat{\theta}) \}$ is the variance estimator of the estimated total, $\widehat{U}(\theta)$, of the $u_k(\theta)$'s evaluated at $\theta = \hat{\theta}$. For example, if $u_k(\theta) = y_k - \theta x_k$ then $\hat{\theta} = \sum_{k \in s_0} w_k y_k / \sum_{k \in s_0} w_k x_k =$

$\widehat{Y}/\widehat{X} = \widehat{R}$ is the estimated ratio, and (3.2.4) reduces to the usual formula

$$\widehat{V}_L(\widehat{\theta}) = \frac{1}{\widehat{X}^2} \widehat{V} \left(\sum_{k \in s_0} w_k u_k(\widehat{\theta}) \right). \quad (3.2.5)$$

Binder (1983) also showed that the statistic $n^{\frac{1}{2}}(\widehat{\theta} - \theta)$ is asymptotically distributed as normal with mean 0 and covariance matrix estimated by $n\widehat{V}_L(\widehat{\theta})$, where $\widehat{V}_L(\widehat{\theta})$ is given by (3.2.4).

3.2.2 Separate estimating equations

For the j -th ($j = 1, 2, \dots, g$) subsample of size m , the separate estimating equations (SEE) are given by

$$\widehat{U}_j^*(\theta) = \frac{N}{m} \sum_{k \in s_j^*} u_k(\theta) = 0; \quad j = 1, \dots, g. \quad (3.2.6)$$

The separate estimating equations estimators $\widehat{\theta}_j^*, j = 1, \dots, g$ are obtained by solving the g equations (3.2.6). In general, we require g separate iterative solutions to get $\widehat{\theta}_1^*, \dots, \widehat{\theta}_g^*$. The inverse-sampling estimator of θ is then given by

$$\widehat{\theta}_g = \frac{1}{g} \sum_{j=1}^g \widehat{\theta}_j^*. \quad (3.2.7)$$

It follows from (3.2.7) that $\widehat{\theta}_\infty = E(\widehat{\theta}_1^* | s_0)$ and $E(\widehat{\theta}_\infty) = E(\widehat{\theta}_1^*)$. Assuming first moment matching with SRS, it follows from (3.2.6) that the bias $E(\widehat{\theta}_1^*) - \theta$ is of order m^{-1} , where m is the subsample size.

The inverse-sampling covariance matrix of $\widehat{\theta}_g$ is given by

$$\widehat{V}_g = \frac{1}{g} \sum_{j=1}^g \widehat{V}_j^* - \frac{1}{g} \sum_{j=1}^g (\widehat{\theta}_j^* - \widehat{\theta}_g)(\widehat{\theta}_j^* - \widehat{\theta}_g)', \quad (3.2.8)$$

where $\widehat{\mathbf{V}}_j^*$ is given by

$$\widehat{\mathbf{V}}_j^* = \widehat{\mathbf{V}}^*(\hat{\boldsymbol{\theta}}_j^*) = \left\{ \frac{\partial \widehat{\mathbf{U}}_j^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j^*} \right\}^{-1} \widehat{\mathbf{V}} \{ \widehat{\mathbf{U}}_j^*(\hat{\boldsymbol{\theta}}_j^*) \} \left\{ \frac{\partial \widehat{\mathbf{U}}_j^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_j^*} \right\}^{-1},$$

$\widehat{\mathbf{V}} \{ \widehat{\mathbf{U}}_j^*(\hat{\boldsymbol{\theta}}_j^*) \}$ is the variance estimator of the j -th subsample total $\widehat{\mathbf{U}}_j^*(\boldsymbol{\theta})$, denoted $\widetilde{\mathbf{V}}_{jS}^*$ (see (3.2.13)) evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_j^*$.

The separate inverse-sampling estimator $\hat{\boldsymbol{\theta}}_g$, as noted earlier, has bias of order m^{-1} and since m may be very much smaller than the original sample size, this bias can be appreciable. Also the inverse-sampling variance estimator (for nonlinear $\hat{\boldsymbol{\theta}}_g$) relies on m being large. Therefore, for small subsample sizes, m , the use of $\hat{\boldsymbol{\theta}}_g$ and $\widehat{\mathbf{V}}_g$ for inferences on $\boldsymbol{\theta}$ may have serious limitations.

3.2.3 Combined estimating equation

To overcome some of the limitations of the separate estimating equations approach, we propose an alternative method: combined estimating equation (CEE). We simply combine the g equations in (3.2.6) before solving for $\boldsymbol{\theta}$. This leads to

$$\widehat{\mathbf{U}}_{gc}(\boldsymbol{\theta}) = \frac{1}{g} \sum_{j=1}^g \widehat{\mathbf{U}}_j^*(\boldsymbol{\theta}) = \frac{1}{g} \sum_{j=1}^g \frac{N}{m} \sum_{k \in s_j^*} \mathbf{u}_k(\boldsymbol{\theta}) = \mathbf{0}. \quad (3.2.9)$$

The solution to (3.2.9), $\hat{\boldsymbol{\theta}}_{gc}$, is the CEE estimator of $\boldsymbol{\theta}$. This estimator leads to asymptotically valid inferences regardless of the subsample size m .

In general, we solve (3.2.9) using the N-R iterations (3.2.3) with $\widehat{\mathbf{U}}(\hat{\boldsymbol{\theta}}^{(r-1)})$ changed to $\widehat{\mathbf{U}}_{gc}(\hat{\boldsymbol{\theta}}^{(r-1)})$ and $\frac{\partial \widehat{\mathbf{U}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(r-1)}}$ changed to

$$\frac{\partial \widehat{\mathbf{U}}_{gc}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \frac{1}{g} \sum_{j=1}^g \frac{N}{m} \sum_{k \in s_j^*} \frac{\partial \mathbf{u}_k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}, \quad (3.2.10)$$

evaluated at $\theta = \hat{\theta}_{gc}$. At convergence, we get the estimator $\hat{\theta}_{gc}$ as well as the observed information matrix given by negating the expression in (3.2.10) evaluated at $\theta = \hat{\theta}_{gc}$. Note that we solve the combined estimating equation only once to get $\hat{\theta}_{gc}$, unlike the separate estimating equations method that solves the g equations (3.2.6) to get $\hat{\theta}_1^*, \dots, \hat{\theta}_g^*$ and then averages the results to obtain $\hat{\theta}_g = \sum_{j=1}^g \hat{\theta}_j^* / g$.

To illustrate the proposed method, consider the special case of ratio $\theta_N = R$, in which case $u_k(\theta) = y_k - \theta x_k$. The solution $\hat{\theta}_{gc} = \hat{R}_{gc}$ is then given by

$$\hat{\theta}_{gc} = \hat{R}_{gc} = \frac{g^{-1} \sum_{j=1}^g \frac{N}{m} \sum_{k \in s_j^*} y_k}{g^{-1} \sum_{j=1}^g \frac{N}{m} \sum_{k \in s_j^*} x_k} = \frac{\hat{Y}_{gc}}{\hat{X}_{gc}},$$

a combined ratio estimator of R .

Theorem 3.1 establishes conditions for which the CEE estimator, $\hat{\theta}_{gc}$ is consistent, Theorem 3.2 establishes consistency of the associated inverse-sampling variance estimator, $\hat{V}_L(\hat{\theta}_{gc})$ and Theorem 3.3 provides asymptotic normality for $\hat{\theta}_{gc}$. These results are presented for univariate θ . We assume the following regularity conditions.

- C1. $\hat{U}(\theta)$ is a continuous function such that $\hat{U}(\hat{\theta}) = 0$.
- C2. $\lim_{n \rightarrow \infty} \frac{1}{N} \hat{U}(\theta) = \Omega(\theta)$, where $\Omega(\theta)$ exists and it is a one-to-one function so that $\Omega^{-1}(\cdot)$ exist.
- C3. $\hat{U}(\cdot)$ is totally differentiable in a neighbourhood of θ_0 and

$$\lim_{\theta \rightarrow \theta_0; n \rightarrow \infty} \left\{ \frac{1}{N} \frac{\partial \hat{U}(\theta)}{\partial \theta} \right\} = \left. \frac{\partial \Omega(\theta)}{\partial \theta} \right|_{\theta = \theta_0} = J(\theta_0) \neq 0.$$

- C4. $\max_i |U_i(\theta^*) - U_i(\hat{\theta})| \leq \epsilon$ if $|\theta^* - \hat{\theta}| \leq \delta(\epsilon)$ for some $\delta(\epsilon) > 0$

- C5. $\max_i \left| \frac{\partial U_i(\theta^*)}{\partial \theta} - \frac{\partial U_i(\hat{\theta})}{\partial \theta} \right| \leq \delta(\epsilon)$ for some $\delta(\epsilon) > 0$
- C6. $\frac{\sqrt{n}}{N} \{ \hat{U}(\theta) - U(\theta) \} \rightarrow_d N(0, \sigma^2)$ for all θ in a neighbourhood of θ_0 .
- C7. $\frac{1}{N} \sum_{i \in s_0} \frac{1}{\pi_i} \left| \frac{\partial U_i(\theta)}{\partial \theta} \right| < \infty$ for all θ in the neighbourhood of θ_0 .
- C8. $\frac{1}{N} \sum_{i \in s_0} \frac{1}{\pi_i} |U_i(\theta)|^2 < \infty$ for all θ in the neighbourhood of θ_0 .

Theorem 3.1 Suppose the full-sample estimator, $\hat{\theta}$ of θ is a unique solution to $\hat{U}(\theta) = 0$, where $\hat{U}(\theta)$ is the Horvitz-Thompson (H-T) estimator of $U(\theta)$. Let $\hat{\theta}_{gc}$ be a combined estimating equation estimator of θ . Then under conditions C1 and C2, $\hat{\theta}_{gc}$ converges in probability to $\hat{\theta}$ as $g \rightarrow \infty$.

Proof

Since $\hat{U}(\hat{\theta}) = 0$ then for any small enough $\epsilon > 0$, $\hat{U}(\hat{\theta} - \epsilon)\hat{U}(\hat{\theta} + \epsilon) < 0$. Under condition C2 and the law of large numbers,

$$\hat{U}_{gc}(\hat{\theta} - \epsilon)\hat{U}_{gc}(\hat{\theta} + \epsilon) \rightarrow_p \hat{U}(\hat{\theta} - \epsilon)\hat{U}(\hat{\theta} + \epsilon) \text{ as } g \rightarrow \infty.$$

Hence for sufficiently large g , $\hat{U}_{gc}(\hat{\theta} - \epsilon)\hat{U}_{gc}(\hat{\theta} + \epsilon) < 0$. Therefore since $\hat{U}_{gc}(\cdot)$ is continuous, there exists $\hat{\theta}_{gc}^* \in (\hat{\theta} - \epsilon, \hat{\theta} + \epsilon)$ such that $\hat{U}_{gc}(\hat{\theta}_{gc}^*) = 0$, i.e., $\hat{\theta}_{gc}^* \rightarrow_p \hat{\theta}$ as $g \rightarrow \infty$. Hence the result. ■

A linearization estimator of $\hat{\theta}_{gc}$ is given by

$$\widehat{V}_L(\hat{\theta}_{gc}) = \left\{ \frac{\partial \widehat{U}_{gc}(\theta)}{\partial \theta'} \Big|_{\theta = \hat{\theta}_{gc}} \right\}^{-1} \widehat{V} \{ \widehat{U}_{gc}(\hat{\theta}_{gc}) \} \left\{ \frac{\partial \widehat{U}_{gc}(\theta)}{\partial \theta'} \Big|_{\theta = \hat{\theta}_{gc}} \right\}^{-1}, \quad (3.2.11)$$

where $\widehat{\mathbf{V}} \{ \widehat{\mathbf{U}}_{gc}(\hat{\boldsymbol{\theta}}_{gc}) \}$ is the variance estimator of the estimated total, $\widehat{\mathbf{U}}_{gc}(\boldsymbol{\theta})$, of the $\mathbf{u}_k(\boldsymbol{\theta})$'s evaluated at $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{gc}$. Since $\widehat{\mathbf{U}}_{gc}(\boldsymbol{\theta})$ is the inverse-sampling estimator of the total $\mathbf{U}(\boldsymbol{\theta})$, it follows that the inverse-sampling covariance matrix of $\widehat{\mathbf{U}}_{gc}(\boldsymbol{\theta})$ is given by

$$\widetilde{\mathbf{V}}_{gS} = \frac{1}{g} \sum_{j=1}^g \widetilde{\mathbf{V}}_{jS}^* - \frac{1}{g} \sum_{j=1}^g \{ \widehat{\mathbf{U}}_j^*(\boldsymbol{\theta}) - \widehat{\mathbf{U}}_{gc}(\boldsymbol{\theta}) \} \{ \widehat{\mathbf{U}}_j^*(\boldsymbol{\theta}) - \widehat{\mathbf{U}}_{gc}(\boldsymbol{\theta}) \}' , \quad (3.2.12)$$

where $\widetilde{\mathbf{V}}_{jS}^*$ is the SRS variance estimator from the j -th subsample, assuming second moment matching. If the matching is with respect to SRS without replacement, then

$$\widetilde{\mathbf{V}}_{jS}^* = \frac{N^2}{m} \left(1 - \frac{m}{N} \right) \frac{1}{m-1} \sum_{k \in s_j^*} \left\{ \mathbf{u}_k(\boldsymbol{\theta}) - \frac{1}{m} \sum_{k \in s_j^*} \mathbf{u}_k(\boldsymbol{\theta}) \right\} \left\{ \mathbf{u}_k(\boldsymbol{\theta}) - \frac{1}{m} \sum_{k \in s_j^*} \mathbf{u}_k(\boldsymbol{\theta}) \right\}' . \quad (3.2.13)$$

In the case of matching to SRS with replacement, we replace $(1 - m/N)$ by 1 in (3.2.13).

Now substituting $\hat{\boldsymbol{\theta}}_{gc}$ for $\boldsymbol{\theta}$ in (3.2.12) we get the inverse-sampling covariance matrix $\widehat{\mathbf{V}}_{gS}$ given by

$$\widehat{\mathbf{V}}_{gS} = \frac{1}{g} \sum_{j=1}^g \widehat{\mathbf{V}}_{jS}^* - \frac{1}{g} \sum_{j=1}^g \widehat{\mathbf{U}}_j^*(\hat{\boldsymbol{\theta}}_{gc}) \widehat{\mathbf{U}}_j^*(\hat{\boldsymbol{\theta}}_{gc})' , \quad (3.2.14)$$

where $\widehat{\mathbf{V}}_{jS}^*$ is obtained from (3.2.13) by substituting $\hat{\boldsymbol{\theta}}_{gc}$ for $\boldsymbol{\theta}$. Note that $\widehat{\mathbf{U}}_{gc}(\hat{\boldsymbol{\theta}}_{gc}) = 0$.

The linearization inverse-sampling variance estimator, for a given g , is now given by

$$\widehat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_{gc}) = \left\{ \frac{\partial \widehat{\mathbf{U}}_{gc}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{gc}} \right\}^{-1} \widehat{\mathbf{V}}_{gS} \left\{ \frac{\partial \widehat{\mathbf{U}}_{gc}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{gc}} \right\}^{-1} . \quad (3.2.15)$$

Under second moment matching with SRS, it is easy to verify that the variance estimator (3.2.15) as $g \rightarrow \infty$ equals the Binder's variance estimator $\widehat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}})$ given by (3.2.4). Thus, the linearization inverse-sampling variance estimator $\widehat{\mathbf{V}}_L(\hat{\boldsymbol{\theta}}_{gc})$ provides

valid inferences on θ for large number of subsamples, g , regardless of the subsample size, m . This result is summarized in Theorem 3.2 for univariate θ .

Theorem 3.2 *Under conditions C1 to C5, and Theorem 3.1, $\hat{V}_L(\hat{\theta}_{gc}) \rightarrow_p \hat{V}_L(\hat{\theta})$ as $g \rightarrow \infty$.*

Proof

The proof follows by showing that

$$J_{gc}(\hat{\theta}_{gc}) \rightarrow_p J(\hat{\theta}) \text{ as } g \rightarrow \infty \quad (3.2.16)$$

and

$$\hat{V} \left\{ \frac{1}{N} \hat{U}_{gc}(\hat{\theta}_{gc}) \right\} - \hat{V} \left\{ \frac{1}{N} \hat{U}(\hat{\theta}) \right\} \rightarrow_p 0 \text{ as } g \rightarrow \infty, \quad (3.2.17)$$

where $J_{gc}(\hat{\theta}_{gc})$ is $J_{gc}(\theta) = \partial U_{gc}(\theta)/(N\partial\theta)$ evaluated at $\theta = \hat{\theta}_{gc}$ and $J(\hat{\theta})$ is $J(\theta) = \partial \hat{U}(\theta)/(N\partial\theta)$ evaluated at $\theta = \hat{\theta}$.

Proof of (3.2.16)

First write $J_{gc}(\hat{\theta}_{gc}) - J(\hat{\theta}) = \{J_{gc}(\hat{\theta}_{gc}) - J_{gc}(\hat{\theta})\} + \{J_{gc}(\hat{\theta}) - J(\hat{\theta})\}$. From Theorem 3.1, $\hat{\theta}_{gc} \rightarrow_p \hat{\theta}$ as $g \rightarrow \infty$ and condition C5, $J_{gc}(\hat{\theta}_{gc}) - J_{gc}(\hat{\theta}) \rightarrow_p 0$ as $g \rightarrow \infty$. By condition C3, and the law of large numbers, $J_{gc}(\hat{\theta}) \rightarrow_p J(\hat{\theta})$ as $g \rightarrow \infty$. Hence (3.2.16) follows.

Proof of (3.2.17)

Rewrite

$$\begin{aligned} \hat{V} \left\{ \frac{1}{N} \hat{U}_{gc}(\hat{\theta}_{gc}) \right\} &= \frac{1}{N^2} \frac{1}{g} \sum_{j=1}^g \hat{V}_{j, \hat{\theta}_{gc}}^* - \frac{1}{N^2} \frac{1}{g} \sum_{j=1}^g \left\{ \hat{U}_j^*(\hat{\theta}_{gc}) \right\}^2 \\ &= \frac{1}{N^2} \frac{1}{g} \sum_{j=1}^g \hat{V}_{j, \hat{\theta}}^* - \frac{1}{N^2} \frac{1}{g} \sum_{j=1}^g \left\{ \hat{U}_j^*(\hat{\theta}) - \bar{U}_{gc}(\hat{\theta}) \right\}^2 + R_1 + R_2 + R_3, \end{aligned}$$

where

$$\hat{V}_{j,\hat{\theta}}^* = \frac{N^2}{m} \left(1 - \frac{m}{N}\right) \frac{1}{m-1} \sum_{i \in s_j^*} \left\{ U_i(\hat{\theta}) - \frac{1}{m} \sum_{i \in s_j^*} U_i(\hat{\theta}) \right\}^2,$$

$$\bar{U}_{gc}(\hat{\theta}) = \frac{1}{g} \sum_{j=1}^g \hat{U}_j^*(\hat{\theta}),$$

$$R_1 = \left(1 - \frac{m}{N}\right) \frac{1}{m(m-1)} \frac{1}{g} \sum_{j=1}^g \sum_{i \in s_j^*} (\epsilon_i - \bar{\epsilon}_j^*)^2,$$

$$R_2 = \left(1 - \frac{m}{N}\right) \frac{1}{m(m-1)} \frac{1}{g} \sum_{j=1}^g \sum_{i \in s_j^*} \left\{ \left[(\epsilon_i - \bar{\epsilon}_j^*)^2 (U_i(\hat{\theta}) - \bar{U}_j^*(\hat{\theta})) \right] \right\},$$

$$R_3 = \left(1 - \frac{m}{N}\right) \frac{1}{m(m-1)} \frac{1}{g} \sum_{j=1}^g \sum_{i \in s_j^*} (\bar{\epsilon}_j^* - \bar{\epsilon}_{gc})^2,$$

$\epsilon_i = U_i(\hat{\theta}_{gc}) - U_i(\hat{\theta})$, $\bar{\epsilon}_j^* = m^{-1} \sum_{i \in s_j^*} \epsilon_i$ and $\bar{\epsilon}_{gc} = g^{-1} \sum_{j=1}^g \bar{\epsilon}_j^*$. Since the first two terms converge to $\hat{V} \{ \hat{U}(\hat{\theta}) \}$, it is sufficient to show that $R_1 \rightarrow_p 0$, $R_2 \rightarrow_p 0$ and $R_3 \rightarrow_p 0$ as $g \rightarrow \infty$: From $(\epsilon_i - \bar{\epsilon}_j^*)^2 \leq 2(\epsilon_i^2 + \bar{\epsilon}_j^{*2}) \leq 2(\epsilon_i^2 + m^{-1} \sum_{i \in s_j^*} \epsilon_i^2) \leq 4\epsilon^2$ by condition C4. Now taking ϵ as small as possible $R_1 \rightarrow 0$, $R_2 \rightarrow 0$ as $g \rightarrow \infty$. Also $R_3 \rightarrow 0$ by noting that $(\bar{\epsilon}_j^* - \bar{\epsilon}_{gc})^2 \leq 2(\bar{\epsilon}_j^{*2} + \bar{\epsilon}_{gc}^2) \leq 2(m^{-1} \sum_{i \in s_j^*} \epsilon_i^2 + (gm)^{-1} \sum_{j=1}^g \sum_{i \in s_j^*} \epsilon_i^2) \leq 4\epsilon^2$ by condition C4 and taking ϵ as small as possible. ■

Theorem 3.3 *Under conditions C1 to C8, $\sqrt{n}(\hat{\theta}_{gc} - \theta_0)$ is asymptotically normal with mean 0 and variance $\sigma^2/J(\theta_0)^2$ as $g \rightarrow \infty$ and $n \rightarrow \infty$.*

Proof

By conditions C1 to C3 and Taylor's expansion of $\widehat{U}(\hat{\theta})$ around $\theta = \theta_0$,

$$-J(\theta_0)\sqrt{n}(\hat{\theta} - \theta_0) = \frac{\sqrt{n}}{N}\widehat{U}(\theta_0) + o_p(1). \quad (3.2.18)$$

By conditions C3, C7 and Taylor's expansion of $\widehat{U}_{gc}(\hat{\theta}_{gc})$ around $\theta = \theta_0$ we have

$$-J(\theta_0)\sqrt{n}(\hat{\theta}_{gc} - \theta_0) = \frac{\sqrt{n}}{\sqrt{g}} \frac{\sqrt{g}}{N} \{\widehat{U}_{gc}(\theta_0) - \widehat{U}(\theta_0)\} + \frac{\sqrt{n}}{N}\widehat{U}(\theta_0) + o_p(1). \quad (3.2.19)$$

Since $\widehat{U}_1^*(\theta), \dots, \widehat{U}_g^*(\theta)$ are conditionally i.i.d with $E\{\widehat{U}_1^*(\theta)|s_0\} = \widehat{U}(\theta) < \infty$ and $V\{\widehat{U}_1^*(\theta)|s_0\} \leq E\{|\widehat{U}_1^*(\theta)|^2\} < \infty$ by condition C8. Then

$$\frac{\sqrt{g}}{N} \{\widehat{U}_{gc}(\theta) - \widehat{U}(\theta)\} \rightarrow_{d|s_0} N(0, \sigma_1^2), \text{ as } g \rightarrow \infty$$

for all θ in the neighbourhood of θ_0 , where σ_1^2 is the limiting variance of $V\{\frac{\sqrt{g}}{N}\widehat{U}_{gc}(\theta)|s_0\}$.

Therefore

$$-J(\theta_0)\sqrt{n}(\hat{\theta}_{gc} - \theta_0) = \frac{\sqrt{n}}{N}\widehat{U}(\theta_0) + o_p(1). \quad (3.2.20)$$

By Lemma 1 of Binder (1983), $n^{\frac{1}{2}}(\hat{\theta} - \theta_0)$ has the same asymptotic distribution as $(-J^{-1}(\theta_0)\widehat{U}(\theta_0)\sqrt{n})/N$. Also by Corollary of Binder (1983), $n^{\frac{1}{2}}(\hat{\theta} - \theta_0)$ is normally distributed with mean 0 and variance $\sigma^2/J^2(\theta_0)$. With these results, assumption C6 and (3.2.20) the proof of the theorem follows. ■

To illustrate the calculation of the linearization inverse-sampling variance estimator $\widehat{V}_L(\hat{\theta}_{gc})$, given by (3.2.15), consider the special case of a ratio $\theta_N = R$ with $u_k(\theta) = y_k - \theta x_k$. We have

$$\tilde{V}_{js}^* = \frac{N^2}{m} \left(1 - \frac{m}{N}\right) \frac{1}{m-1} \sum_{k \in s_j^*} \{u_k(\theta) - \bar{u}_j^*(\theta)\}^2, \quad (3.2.21)$$

where $\bar{u}_j^*(\theta) = \bar{y}_j^* - \theta \bar{x}_j^*$ and $(\bar{y}_j^*, \bar{x}_j^*)$ are the j -th subsample means. Further,

$$-\frac{\partial \widehat{U}_{gc}(\theta)}{\partial \theta} = \frac{N}{g} \sum_{j=1}^g \bar{x}_j^* = \widehat{X}_{gc} \quad (3.2.22)$$

and

$$\widehat{U}_j^*(\theta) = N(\bar{y}_j^* - \theta \bar{x}_j^*). \quad (3.2.23)$$

It now follows from (3.2.21)-(3.2.23) that the CEE-based linearization variance estimator (3.2.15) is identical to the linearization variance estimator (3.2.5).

Turning to linear regression with $\mathbf{u}_k(\theta) = \mathbf{x}_k(y_k - \mathbf{x}_k' \theta)$, we have

$$\widetilde{\mathbf{V}}_{jS}^* = \frac{N^2}{m} \left(1 - \frac{m}{N}\right) \frac{1}{m-1} \sum_{k \in s_j^*} \{\mathbf{u}_k(\theta) - \bar{\mathbf{u}}_j^*(\theta)\} \{\mathbf{u}_k(\theta) - \bar{\mathbf{u}}_j^*(\theta)\}', \quad (3.2.24)$$

where $\bar{\mathbf{u}}_j^*(\theta) = 1/m \sum_{k \in s_j^*} \mathbf{x}_k(y_k - \mathbf{x}_k' \theta)$. Also,

$$-\frac{\partial \widehat{U}_{gc}(\theta)}{\partial \theta'} = \frac{N}{g} \sum_{j=1}^g \frac{1}{m} \sum_{k \in s_j^*} \mathbf{x}_k \mathbf{x}_k',$$

and

$$\widehat{U}_j^*(\theta) = \frac{N}{m} \sum_{k \in s_j^*} \mathbf{x}_k(y_k - \mathbf{x}_k' \theta).$$

Finally consider the case of logistic regression with $\mathbf{u}_k(\theta) = \mathbf{x}_k(y_k - \mu(\theta))$. In this case, $\widetilde{\mathbf{V}}_{jU}^*$ is given by (3.2.24) with $\mathbf{u}_k(\theta) = \mathbf{x}_k(y_k - \mu(\theta))$. Also

$$-\frac{\partial \widehat{U}_{gc}(\theta)}{\partial \theta'} = \frac{N}{g} \sum_{j=1}^g \frac{1}{m} \sum_{k \in s_j^*} \mu_k(\theta)(1 - \mu_k(\theta)) \mathbf{x}_k \mathbf{x}_k',$$

and

$$\widehat{U}_j^*(\theta) = \frac{N}{m} \sum_{k \in s_j^*} \mathbf{x}_k(y_k - \mu_k(\theta)).$$

It is important to note again that the estimator $\hat{\theta}_{gc}$ and the associated variance estimator $\widehat{V}_L(\hat{\theta}_{gc})$ can be implemented from a microdata file providing g subsamples, each of size m . Neither the survey weights w_k nor the cluster identifiers are needed so that confidentiality of microdata may be preserved.

3.3 Poststratification Estimation

In this section, we study an application of inverse sampling to poststratification estimation. Poststratification information is commonly used to improve the precision of estimators of finite population parameters as well as to ensure benchmarking to known marginal totals associated with the categories of the poststratification variables.

In Subsection 3.3.1, we introduce a full-sample estimator of a total using poststratification information, followed by an inverse-sampling approach, combined poststratification estimation method that we propose.

3.3.1 Estimation of a Total

(i) Full-sample poststratification estimation

Let $\mathbf{x}_k$ be the k -th unit of a $p \times 1$ vector of indicator variables corresponding to the categories of poststratification variables with known marginal totals, $\mathbf{X} = \sum_{k \in U} \mathbf{x}_k$. To ensure benchmarking, the design weights, $w_k, k \in s_0$ (in this thesis, we restrict $w_k = 1/\pi_k$) are adjusted to construct calibration weights, $\tilde{w}_k(s_0)$ satisfying $\sum_{k \in s_0} \tilde{w}_k(s_0) \mathbf{x}_k = \mathbf{X}$. The summation, $\sum_{k \in s_0}$ is over the full sample, s_0 . The Generalized regression (GREG) is commonly used to construct the calibration weights.

A GREG estimator of a total $Y = \sum_{k \in U} y_k$ is given by

$$\hat{Y}_r = \hat{Y}_r(y) = \mathbf{X}' \hat{\mathbf{B}} = \sum_{k \in U} \hat{y}_k, \quad (3.3.25)$$

where $\hat{y}_k = \mathbf{x}'_k \hat{\mathbf{B}}$ is the predictor of y_k ,

$$\hat{\mathbf{B}} = \left(\sum_{k \in s_0} \frac{\mathbf{x}_k \mathbf{x}'_k}{\pi_k} \right)^{-1} \left(\sum_{k \in s_0} \frac{\mathbf{x}_k y_k}{\pi_k} \right). \quad (3.3.26)$$

Substituting (3.3.26) in (3.3.25) and simplifying to obtain

$$\hat{Y}_r = \sum_{k \in s_0} \frac{a_k(s_0)}{\pi_k} y_k = \sum_{k \in s_0} \tilde{w}_k(s_0) y_k, \quad (3.3.27)$$

where

$$a_k(s_0) = \mathbf{X}' \left(\sum_{k \in s_0} \frac{\mathbf{x}_k \mathbf{x}'_k}{\pi_k} \right)^{-1} \mathbf{x}_k, \quad (3.3.28)$$

is the adjustment factor also referred to as g-weight (see Särndal, Swensson and Wretman, 1992, Chapter 6). We observed in (3.3.27) that the calibration weight $\tilde{w}_k(s_0)$ is a product of a poststratification adjustment factor $a_k(s_0)$ and the design weight. We note that $\hat{Y}_r(\mathbf{x}) = \sum_{k \in s_0} a_k(s_0) \mathbf{x}_k / \pi_k = \mathbf{X}'$.

A linearization variance estimator of $\hat{Y}_r$ is given by (see Särndal, Swensson and Wretman, 1992 Chapter 6).

$$\hat{V}(\hat{Y}_r) = \sum_{k,l} \sum_{k,l \in s_0} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} \{a_k(s_0) e_k\} \{a_l(s_0) e_l\}, \quad (3.3.29)$$

where $e_k = y_k - \mathbf{x}'_k \hat{\mathbf{B}}$.

A computationally simpler variance estimator of $\hat{Y}_r$ is obtained by replacing $a_k(s_0) e_k$ by e_k in (3.3.29) to obtain

$$\hat{V}(\hat{Y}_r) = \sum_{k,l} \sum_{k,l \in s_0} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} e_k e_l. \quad (3.3.30)$$

As discussed in Särndal, Swensson and Wretman 1992, the two variance estimators (3.3.29) and (3.3.30) provide approximately good $100(1 - \alpha)\%$ coverage rate in repeated large samples. However, (3.3.29) is preferred in some cases.

We note that estimation of the full-sample variance estimators may require identification of strata or clusters which in general may not be available on a public use microdata file due to the need to protect privacy of respondents. We propose an alternative approach, combined poststratification estimation, where the issue of stratum or cluster confidentiality may not be a problem.

(ii) Combined poststratification estimation

In this method, we combine the g inverse-samples, $s_j^* = \{(y_k, \mathbf{x}_k); k = 1, \dots, m\}; j = 1, \dots, g$ to obtain the combined estimating equation estimator, $\widehat{B}_{gc}$ of the census regression parameter B . The combined generalized regression estimator of Y is then given by

$$\widehat{Y}_{r,gc} = \widehat{Y}_{r,gc}(y) = \mathbf{X}' \widehat{B}_{gc} = \frac{N}{gm} \sum_{j=1}^g \sum_{k \in s_j^*} a_k(s^*) y_k, \quad (3.3.31)$$

where

$$\widehat{B}_{gc} = \left(\frac{N}{gm} \sum_{j=1}^g \sum_{k \in s_j^*} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \frac{N}{gm} \sum_{j=1}^g \sum_{k \in s_j^*} \mathbf{x}_k y_k = \widehat{T}_{gxx}^{-1} \widehat{t}_{gxy}, \text{ say}$$

and

$$a_k(s^*) = \mathbf{X}' \left(\frac{N}{gm} \sum_{j=1}^g \sum_{k \in s_j^*} \mathbf{x}_k \mathbf{x}_k \right)^{-1} \mathbf{x}_k.$$

We note that

$$\widehat{Y}_{r,gc}(\mathbf{x}) = \frac{N}{gm} \sum_{j=1}^g \sum_{k \in s_j^*} a_k(s^*) \mathbf{x}_k = \mathbf{X}'.$$

The following theorem shows that for large number of repeated subsamples, the combined generalized regression estimator preserves the full-sample GREG estimator.

Theorem 3.4 *Under mild regularity conditions, the combined generalized regression estimator, $\hat{Y}_{r,gc}$ converges almost surely to the full-sample generalized regression estimator, $\hat{Y}_r$ for large number of repeated subsamples, g regardless of the subsample size, m .*

Proof

We first note that

$$\hat{T}_{gxx} = \frac{N}{gm} \sum_{j=1}^g \sum_{k \in s_j^*} \mathbf{x}_k \mathbf{x}_k'$$

converges almost surely to

$$E \left(\frac{N}{m} \sum_{k \in s_1^*} \mathbf{x}_k \mathbf{x}_k' | s_0 \right) = \sum_{k \in s_0} \frac{\mathbf{x}_k \mathbf{x}_k'}{\pi_k} = \hat{T}_{xx}, \text{ say, as } g \rightarrow \infty.$$

Similarly,

$$\hat{t}_{gxy} = \frac{N}{gm} \sum_{j=1}^g \sum_{k \in s_j^*} \mathbf{x}_k y_k$$

converges almost surely to

$$E \left(\frac{N}{m} \sum_{k \in s_1^*} \mathbf{x}_k y_k | s_0 \right) = \sum_{k \in s_0} \frac{\mathbf{x}_k y_k}{\pi_k} = \hat{t}_{xy}, \text{ say, as } g \rightarrow \infty.$$

Therefore using standard results, $\hat{B}_{gc} = \hat{T}_{gxx}^{-1} \hat{t}_{gxy}$ converges almost surely to $\hat{B} = \hat{T}_{xx}^{-1} \hat{t}_{xy}$ as $g \rightarrow \infty$. Hence the result. ■

A linearization inverse-sampling variance estimator of $\hat{Y}_{r,gc}$ is obtained by first establishing the following: We first note that, $\hat{Y}_{r,gc}$ may be rewritten as $\hat{Y}_{r,gc} = \hat{Y}_g + (X - \hat{X}_g)' \hat{T}_{gxx}^{-1} \hat{t}_{gxy} = f(\hat{Y}_{r,gc}, \hat{X}_g, \hat{T}_{gxx}, \hat{t}_{gxy})$, where $\hat{Y}_g = \{N/(gm)\} \sum_{j=1}^g \sum_{i \in s_j^*} y_i$ and $\hat{X}_g = \{N/(gm)\} \sum_{j=1}^g \sum_{i \in s_j^*} \mathbf{x}_i$. A Taylor linearization of $f(\hat{Y}_{r,gc}, \hat{X}_g, \hat{T}_{gxx}, \hat{t}_{gxy})$ around the population quantities $Y, \mathbf{X}, \mathbf{T}$ and $\mathbf{t}$ leads to

$$\begin{aligned} \hat{Y}_{r,gc} - Y &\approx (\hat{Y}_g - \hat{\mathbf{X}}_g' \mathbf{B}) - (Y - \mathbf{X}' \mathbf{B}) = \frac{N}{gm} \sum_{j=1}^g \sum_{k \in s_j^*} (y_k - \mathbf{x}_k' \mathbf{B}) - (Y - \mathbf{X}' \mathbf{B}) \\ &= \frac{N}{gm} \sum_{j=1}^g \sum_{k \in s_j^*} u_k(\mathbf{B}) - (Y - \mathbf{X}' \mathbf{B}) = \frac{1}{g} \sum_{j=1}^g \tilde{U}_{rj}^* - (Y - \mathbf{X}' \mathbf{B}) \\ &= \tilde{U}_{r,gc} - (Y - \mathbf{X}' \mathbf{B}), \end{aligned}$$

where $\tilde{U}_{rj}^* = (N/m) \sum_{k \in s_j^*} u_k(\mathbf{B})$, and $u_k(\mathbf{B}) = y_k - \mathbf{x}_k' \mathbf{B}$. Now using the standard relationship between unconditional and conditional variances we have

$$V(\tilde{U}_{r,gc}) = V(\tilde{U}_{r1}^*) - \frac{g-1}{g} EV(\tilde{U}_{r1}^* | s_0).$$

The inverse-sampling variance of $\tilde{U}_{r,gc}$ is given by

$$\hat{V}(\tilde{U}_{r,gc}) = \frac{1}{g} \sum_{j=1}^g \tilde{V}_{rj}^* - \frac{1}{g} \sum_{j=1}^g (\tilde{U}_{rj}^* - \tilde{U}_{r,gc})^2, \quad (3.3.32)$$

where

$$\tilde{V}_{rj}^* = \frac{N^2}{m} \left(\frac{1}{m} - \frac{1}{N} \right) \frac{1}{m-1} \sum_{k \in s_j^*} (u_k(\mathbf{B}) - \bar{u}_j(\mathbf{B}))^2$$

and $\bar{u}_j(\mathbf{B}) = m^{-1} \sum_{k \in s_j^*} u_k(\mathbf{B})$. Substituting $\hat{\mathbf{B}}_{gc}$ for $\mathbf{B}$ into (3.3.32) leads to the inverse-sampling variance estimator of $\hat{Y}_{r,gc}$ given by

$$\hat{V}(\hat{Y}_{r,gc}) = \frac{1}{g} \sum_{j=1}^g \hat{V}_{rjS}^* - \frac{1}{g} \sum_{j=1}^g \hat{U}_{rj}^{*2}, \quad (3.3.33)$$

where $\hat{U}_{rj}^*$ and $\hat{V}_{rjS}^*$ are $\tilde{U}_{rj}^*$ and $\tilde{V}_{rjS}^*$ evaluated at $B = \hat{B}_{gc}$. Note that $\tilde{U}_{r,gc}$ evaluated at $B = \hat{B}_{gc}$ is zero. Note also that, as g goes to infinity and under mild regularity conditions, the variance estimator of $\hat{Y}_{r,gc}$, $\hat{V}(\hat{Y}_{r,gc})$ converges to the full-sample variance estimator $\hat{V}(\hat{Y}_r)$ given by (3.3.30).

3.3.2 Estimation of Nonlinear Parameters

We extend the results of Subsection 3.3.1 to finite population parameters that are solutions to the census estimating equations. We first introduce results for the full-sample estimators and then propose an inverse-sampling approach, combined poststratification estimating equation method.

(i) Full-sample postratification estimating equations

Suppose θ_N , a vector of finite population parameter is the solution to the census estimating equation given by

$$U(\theta) = \sum_{k \in U} u_k(\theta) = 0, \quad (3.3.34)$$

where $\sum_{k \in U}$ is the summation over the finite population U of size N , and the estimating functions $u_k(\theta)$ are suitably chosen. For example, the choice $u_k(\theta) = y_k - \theta$ leads to a population mean, $\theta = \bar{Y}$, for $u_k(\theta) = y_k - x_k\theta$, we have a population ratio $\theta_N = R = Y/X$, $u_k(\theta) = x_k(y_k - \mu_k(\theta))$ with $\mu_k(\theta) = x_k'\theta$ gives the least squares regression parameter $\theta_N = (\sum_{k \in U} x_k x_k')^{-1} \sum_{k \in U} x_k y_k$ and for $u_k(\theta) = x_k(y_k - \mu_k(\theta))$ with $\mu_k(\theta) = \exp(x_k'\theta) / [1 + \exp(x_k'\theta)]$ results in the logistic regression parameter θ_N .

The generalized regression (GREG) estimator, $\hat{\theta}$, of θ_N is the solution to the

full-sample estimating equation

$$\widehat{U}(\theta) = \sum_{k \in s_0} \tilde{w}_k(s_0) u_k(\theta) = 0, \quad (3.3.35)$$

where $\tilde{w}_k(s_0)$ is as defined in Subsection 3.3.1 and $\widehat{U}(\theta)$ is the total of the $u_k(\theta)$'s.

In general, $\hat{\theta}$ is obtained by solving (3.3.35) iteratively. A linearization variance estimator of $\hat{\theta}$ is given by

$$\widehat{V}_L(\theta) = \left\{ \frac{\partial \widehat{U}(\theta)}{\partial \theta'} \bigg|_{\theta=\hat{\theta}} \right\}^{-1} \widehat{V} \{ \widehat{U}(\hat{\theta}) \} \left\{ \frac{\partial \widehat{U}(\theta)}{\partial \theta'} \bigg|_{\theta=\hat{\theta}} \right\}^{-1}, \quad (3.3.36)$$

where $\widehat{V} \{ \widehat{U}(\hat{\theta}) \}$ is a variance estimator of the estimated total $\widehat{U}(\hat{\theta})$. Hidiroglou, Rao and Yung (1999) showed that

$$\widehat{V}(\widehat{U}(\hat{\theta})) = v\{a_k(s_0)e_k^*(\hat{\theta})\}, \quad (3.3.37)$$

where the j -th element of $e_k^*(\hat{\theta})$ is given by

$$e_{kj}^*(\hat{\theta}) = u_{kj}(\hat{\theta}) - \widehat{B}_j'(\hat{\theta})x_k \quad (3.3.38)$$

and

$$\widehat{B}_j(\hat{\theta}) = \left(\sum_{k \in s_0} w_k x_k x_k' \right)^{-1} \sum_{k \in s_0} w_k x_k u_{kj}(\hat{\theta}). \quad (3.3.39)$$

(ii) Combined postratification estimating equation

A combined postratification estimating equation estimator, $\hat{\theta}_{gc}$ is the solution to the estimating equation given by

$$\widehat{U}_{gc}(\theta) = \frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j^*} \tilde{w}_k(s^*) u_k^*(\theta) = \frac{1}{g} \sum_{j=1}^g \widehat{U}_j^*(\theta) = 0, \quad (3.3.40)$$

where $\tilde{w}_k(s^*) = X' (g^{-1} \sum_{j=1}^g \sum_{k \in s_j^*} x_k x_k')^{-1} x_k$. It is also easy to see that

$g^{-1} \sum_{j=1}^g \sum_{k \in s_j^*} \tilde{w}_k(s^*) x_k' = X'$. A linearization inverse-sampling variance estimator

of $\hat{\theta}_{gc}$ is given by

$$\widehat{V}_{gc} = \left\{ \frac{\partial \widehat{U}_{gc}(\theta)}{\partial \theta'} \bigg|_{\theta=\hat{\theta}_{gc}} \right\}^{-1} \widehat{V} \{ \widehat{U}_{gc}(\hat{\theta}_{gc}) \} \left\{ \frac{\partial \widehat{U}_{gc}(\theta)}{\partial \theta'} \bigg|_{\theta=\hat{\theta}_{gc}} \right\}^{-1}, \quad (3.3.41)$$

where

$$\widehat{V} \{ \widehat{U}_{gc}(\hat{\theta}_{gc}) \} = \frac{1}{g} \sum_{j=1}^g \widehat{V}_{js}^* - \frac{1}{g} \sum_{j=1}^g \widehat{U}_j^*(\theta_{gc}) \widehat{U}_j^*(\theta_{gc})', \quad (3.3.42)$$

$\widehat{U}_j^*(\hat{\theta}_{gc})$ and $\widehat{V}_{js}^*$ are $\widehat{U}_j^*(\theta)$ and $\widehat{V} \{ \widehat{U}_j^*(\theta) \}$ evaluated at $\theta = \hat{\theta}_{gc}$,

$$\widehat{V} \{ \widehat{U}_j^*(\theta) \} = \frac{N^2}{m} \left(1 - \frac{m}{N} \right) \frac{1}{m-1} \sum_{k \in s_j^*} \{ \widetilde{U}_k^*(\theta) - \overline{U}_j^*(\theta) \} \{ \widetilde{U}_k^*(\theta) - \overline{U}_j^*(\theta) \}', \quad (3.3.43)$$

the l -th element of $\widehat{U}_k^*(\theta)$, $\tilde{u}_{kl}^*(\theta) = u_{kl}^*(\theta) - \mathbf{x}_k' \mathbf{B}_l$ is evaluated at $\widehat{B}_l^*(\hat{\theta}_{gc})$ and

$$\widehat{B}_l^*(\hat{\theta}_{gc}) = \left(\frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j^*} \tilde{w}_k(s^*) \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \frac{1}{g} \sum_{j=1}^g \sum_{k \in s_j^*} \tilde{w}_k(s^*) \mathbf{x}_k u_{kl}(\hat{\theta}_{gc}).$$

3.4 Analyses of Categorical Complex Survey Data

Standard methods for analyzing categorical data were developed under the assumption of multinomial sampling. Many researchers in subject matter areas such as the social sciences and the health sciences use standard methods to analyze complex survey categorical data, without taking into account the complex nature of the data, (e.g., stratification and clustering). This may lead to asymptotically invalid statistical inferences.

In recent years many methods have been developed that take account of the survey design features in the analyses of categorical survey data. Nathan (1975), Koch, Freeman and Freeman (1975), Shuster and Dawning (1976) and others developed methods

based on the Wald statistic that account for the design features. Cohen (1976), Altham (1976), Fellegi (1980), and Rao and Scott (1979, 1981, 1984) and others showed that appropriate adjustment on the standard Pearson chi-square statistic lead to asymptotically valid analyses. The above methods depend on the knowledge of certain design effects or estimated covariance matrices of cell estimators.

Generally these methods are tailored to fit the data. Also users of published survey data, such as analysts and researchers, may not have the required design information, such as cluster identifiers and design effects, to implement these methods adequately. In addition, these methods require specialized software applications for their implementation.

An alternative approach we propose is the use of inverse sampling methods. Instead of developing new methods to fit the complex survey data, we use inverse sampling methods to tailor the data to fit standard methods.

In Subsection 3.4.1, we summarize some results of goodness-of-fit test statistic and the Wald test statistic for analyzing complex survey data. We propose and study properties of the inverse-sampling goodness-of-fit tests in Subsection 3.4.2. Limited simulation studies were also done to compare the various methods.

3.4.1 Full-sample Goodness-of-Fit Tests

There are many applications in complex surveys, where goodness-of-fit problems for finite population proportions, p_i , are of interest. For example, one might wish to test the hypothesis that the population proportion, p_i , of a household characteristic in a

household survey is the same as that found in the previous census, p_{0i} .

Consider a finite population $U = \{1, \dots, i, \dots, N\}$ of size N . Suppose U is partitioned into K nonoverlapping and exhaustive categories (cells) U_i ($i = 1, \dots, K$). The size of U_i is denoted by N_i . Let p_i denote the finite population proportion for the i -th category such that $p_i \geq 0$ and $\sum_{i=1}^K p_i = 1$. Consider the goodness-of-fit problem for testing $H_0 : p_i = p_{0i}$ ($i = 1, \dots, K$) for specified p_{0i} .

Let s_0 be a sample of size n drawn from the finite population according to a specified complex design $p_0(s_0)$. A consistent estimator of p_i under $p_0(s_0)$ is given by

$$\hat{p}_i = \frac{\sum_{j \in s_0} y_{ij} / \pi_j}{\sum_{j \in s_0} 1 / \pi_j}, \quad i = 1, \dots, K \quad (3.4.44)$$

where π_j is the inclusion probability of the j -th unit, $y_{ij} = 1$ if $j \in U_i$; 0 otherwise, and $\sum_{i=1}^K \hat{p}_i = 1$ since $\sum_{i=1}^K y_{ij} = 1$.

The generalized Wald statistic for testing H_0 is given by

$$X_W^2 = n(\hat{\mathbf{p}} - \mathbf{p}_0)' \widehat{\mathbf{V}}^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0), \quad (3.4.45)$$

where $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_{K-1})'$, $\mathbf{p}_0 = (p_{01}, \dots, p_{0,K-1})'$ and $\widehat{\mathbf{V}}/n$ is an estimator of the covariance matrix $\mathbf{V}/n$, of $\hat{\mathbf{p}}$. The X_W^2 is asymptotically distributed as a χ_{K-1}^2 under H_0 as $n \rightarrow \infty$.

A Pearson chi-squared statistic for testing H_0 is given by

$$X^2 = n \sum_{i=1}^K \frac{(\hat{p}_i - p_{0i})^2}{p_{0i}} = n(\hat{\mathbf{p}} - \mathbf{p}_0)' \mathbf{P}_0^{-1} (\hat{\mathbf{p}} - \mathbf{p}_0), \quad (3.4.46)$$

where $\mathbf{P}_0 = \text{diag}(\mathbf{p}_0) - \mathbf{p}_0 \mathbf{p}_0'$. We restate the asymptotic distribution of X^2 , given by Rao and Scott (1981), in the following theorem.

Theorem 3.5 Under the null hypothesis $H_0 : p_i = p_{0i} (i = 1, \dots, K)$, X^2 may be written as $\sum_{i=1}^{K-1} \lambda_{0i} z_i^2$, where $z_1, \dots, z_{K-1}$ are asymptotically independent $N(0, 1)$ random variables and λ_{0i} 's are eigenvalues of $\Delta_0 = P_0^{-1}V$ ($\lambda_{01} \geq \lambda_{02} \geq \dots \geq \lambda_{0,K-1}$).

In practice, since λ_{0i} 's or consistent estimators $\hat{\lambda}_i$ are unknown, Rao and Scott (1981) provided simple corrections to the asymptotic distribution of X^2 : (i) first-order correction; (ii) second-order correction.

The first-order corrected statistic proposed by Rao and Scott (1981) treats

$$X_c^2 = \frac{X^2}{\hat{\lambda}} \approx \frac{\sum_{i=1}^{K-1} \lambda_i z_i^2}{\hat{\lambda}} =: Y \quad (3.4.47)$$

as a χ_{K-1}^2 random variable under H_0 , where $\hat{\lambda} = \sum_{i=1}^{K-1} \hat{\lambda}_i / (K-1)$ and $\hat{\lambda}_i$'s are eigenvalues of $\widehat{D} = \widehat{P}^{-1}\widehat{V}$ and $\widehat{P} = \text{diag}(\hat{p}) - \hat{p}\hat{p}'$.

It is important to note that $\hat{\lambda} = (K-1)^{-1} \sum_{i=1}^K (1-\hat{p}_i)\hat{d}_i$, where $\hat{d}_i = \widehat{V}_{ii} / \{\hat{p}_i(1-\hat{p}_i)\}$ is the estimated design effects of the i -th cell and $\widehat{V}_{ii}$ is the i -th diagonal element of $\widehat{V}$. Thus, only cell design effects are needed to implement the first-order correction.

We note that Y has the same expectation as χ_{K-1}^2 but its variance, $V(Y) = 2(K-1) + 2 \sum_{i=1}^{K-1} (\lambda_{0i} - \bar{\lambda}_0)^2 / \bar{\lambda}_0^2$, is larger than $V(\chi_{K-1}^2) = 2(K-1)$ unless all the λ_{0i} 's are equal. Thus, treating X_c^2 as χ_{K-1}^2 under H_0 will tend to give significance levels somewhat larger than the nominal test levels.

The second-order correction statistic is a better approximation provided additional information on $\hat{\lambda}_i$'s is available. For example, if $\sum_{i=1}^{K-1} \hat{\lambda}_i^2$ is known, Rao and Scott (1981) treats

$$X_s^2 = \frac{X_c^2}{1 + \hat{\alpha}^2} \quad (3.4.48)$$

as χ^2_ν , where $\nu = (K-1)/(1+\hat{a}^2)$ and $\hat{a}^2 = \sum_{i=1}^{K-1} (\hat{\lambda}_i - \hat{\bar{\lambda}})^2 / \{(K-1)\hat{\bar{\lambda}}\}$ is the square of coefficient of variation of the $\hat{\lambda}_i$'s. Note that $\sum_{i=1}^{K-1} \hat{\lambda}_i^2 = \sum \sum_{i,j=1}^K \hat{V}_{ij}^2 / (\hat{p}_i \hat{p}_j)$, so that the second-order correction can be implemented if $\hat{V}$ is known, the same information required for the Wald statistic, X_W^2 . However, finite sample studies (Thomas and Rao, 1987) indicate that X_S^2 performs better than X_W^2 in terms of significance levels.

3.4.2 Inverse-sampling Goodness-of-fit Tests

Suppose s_0 is a sample of m clusters drawn from M clusters in a population of N elements according to a one-stage or two-stage cluster sampling design. An inverse sample s^* of size m elements is drawn from s_0 such that the unconditional probability of s^* , $p(s^*)$, matches simple random sampling either exactly or approximately. The exact or approximate inverse sampling schemes for a number of one-stage and two-stage cluster sampling designs are detailed in Subsection 2.2.2 and Subsection 2.2.3.

Suppose a sequence of g independent inverse-samples s_j^* ($j = 1, \dots, g$) each of size m are generated. Then the subsampling estimators, $\hat{p}_j^*$ ($j = 1, \dots, g$) of p are

$$\hat{p}_j^* = \frac{1}{m} \sum_{l \in s_j^*} y_l, \quad j = 1, \dots, g \quad (3.4.49)$$

where $y_l = (y_{l1}, \dots, y_{l,K-1})'$, $y_{li} = 1$, if $l \in U_i$ ($i = 1, \dots, K-1$); 0 otherwise $\hat{p}_j^* = (\hat{p}_{j1}^*, \dots, \hat{p}_{j,K-1}^*)'$ and $\sum_{i=1}^K \hat{p}_{ij}^* = 1$. The inverse-sampling estimator of p is then given by

$$\hat{p}_g = \frac{1}{g} \sum_{j=1}^g \hat{p}_j^* = \frac{1}{gm} \sum_{j=1}^g \sum_{l \in s_j^*} y_l, \quad (3.4.50)$$

and the inverse-sampling covariance matrix of $\hat{\mathbf{p}}_g$ is

$$\widehat{\mathbf{V}}_g = \frac{1}{g} \sum_{j=1}^g \widehat{\mathbf{V}}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\mathbf{p}}_j^* - \hat{\mathbf{p}}_g) (\hat{\mathbf{p}}_j^* - \hat{\mathbf{p}}_g)', \quad (3.4.51)$$

where $\widehat{\mathbf{V}}_j^* = \widehat{\mathbf{V}}(\hat{\mathbf{p}}_j^*) = \frac{1}{m} \{ \text{diag}(\hat{\mathbf{p}}_j^*) - \hat{\mathbf{p}}_j^* \hat{\mathbf{p}}_j^{*'} \}$. The inverse-sampling generalized Wald statistic under H_0 is

$$X_{g,W}^2 = m(\hat{\mathbf{p}}_g - \mathbf{p}_0)' \widehat{\mathbf{V}}_g^{-1} (\hat{\mathbf{p}}_g - \mathbf{p}_0), \quad (3.4.52)$$

and the inverse-sampling Pearson chi-squared statistic is

$$X_g^2 = m(\hat{\mathbf{p}}_g - \mathbf{p}_0)' \mathbf{P}_0^{-1} (\hat{\mathbf{p}}_g - \mathbf{p}_0). \quad (3.4.53)$$

The following theorem establishes the asymptotic distribution of the inverse-sampling generalized Wald statistic, $X_{g,W}^2$.

Theorem 3.6 *Under the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0$, and assume the regularity conditions C1-C4 given in Subsection 2.4.2 for estimation of proportions, then the inverse-sampling generalized Wald statistic $X_{g,W}^2$ converges to a χ_{K-1}^2 as $g \rightarrow \infty$ and then $m \rightarrow \infty$.*

Proof

Apply Theorem 2.7 and Theorem 2.8 to the components of $\hat{\mathbf{p}}_g = (\hat{p}_{g1}, \dots, \hat{p}_{g,K-1})'$, where $\hat{p}_{gi} = (gm)^{-1} \sum_{j=1}^g \sum_{l \in s_j^*} y_{il}$, $i = 1, \dots, K-1$, it follows that $m^{\frac{1}{2}}(\hat{\mathbf{p}}_g - \mathbf{p})$ is asymptotically normally distributed with zero mean and covariance $\mathbf{V}$ as $g \rightarrow \infty$ and $m \rightarrow \infty$. The rest of the proof follows easily from standard results. ■

Theorem 3.7 below provides the corresponding asymptotic distribution of the inverse-sampling Pearson chi-squared statistic.

Theorem 3.7 *Under the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0$, the inverse-sampling Pearson chi-squared statistic, X_g^2 converges in distribution to $\sum_{i=1}^{K-1} \lambda_{0i} Z_i^2$, provided the regularity conditions C1-C4 given in Subsection 2.4.2 hold, where Z_i 's are independent $N(0, 1)$ random variables, λ_{0i} 's are the eigenvalues of $\mathbf{P}_0^{-1} \mathbf{V}$ ($\lambda_{01} \geq \lambda_{02} \geq \dots \geq \lambda_{0,K-1} > 0$).*

Proof

We first note that under H_0 ,

$$\begin{aligned} X_g^2 &= m(\hat{\mathbf{p}}_g - \mathbf{p}_0)' \mathbf{P}_0^{-1} (\hat{\mathbf{p}}_g - \mathbf{p}_0) \\ &= \left\{ m^{\frac{1}{2}} \mathbf{V}^{-\frac{1}{2}} (\hat{\mathbf{p}}_g - \mathbf{p}_0) \right\}' \left(\mathbf{V}^{-\frac{1}{2}} \mathbf{P}_0^{-1} \mathbf{V}^{-\frac{1}{2}} \right) \left\{ m^{\frac{1}{2}} \mathbf{V}^{-\frac{1}{2}} (\hat{\mathbf{p}}_g - \mathbf{p}_0) \right\}. \end{aligned}$$

Apply Theorem 2.7 and Theorem 2.8 to the components of $\hat{\mathbf{p}}_g = (\hat{p}_{g1}, \dots, \hat{p}_{g,K-1})'$, where $\hat{p}_{gi} = (gm)^{-1} \sum_{j=1}^g \sum_{l \in s_j^*} y_{il}$, $i = 1, \dots, K-1$, it follows that $m^{\frac{1}{2}}(\hat{\mathbf{p}}_g - \mathbf{p})$ is asymptotically normally distributed with zero mean and covariance $\mathbf{V}$ as $g \rightarrow \infty$ and $m \rightarrow \infty$. The rest of the proof follows easily from standard results. ■

The inverse-sampling first-order correction to X_g^2 that we propose is given by

$$X_{g,c}^2 = \frac{X_g^2}{\hat{\lambda}_g}, \quad (3.454)$$

where $\hat{\lambda}_g = (K-1)^{-1} \sum_{i=1}^{K-1} \hat{\lambda}_{g,i}$, $\hat{\lambda}_{g,i}$'s are the eigenvalues of $\widehat{\mathbf{P}}_g^{-1} \widehat{\mathbf{V}}_g$. Under the null hypothesis, we treat $X_{g,c}^2$ as asymptotically equivalent to χ_{K-1}^2 .

The inverse-sampling second-order correction we propose is given by

$$X_{g,s}^2 = \frac{X_{g,c}^2}{1 + \hat{a}_g^2} \quad (3.4.55)$$

which is treated as $\chi_{\nu_g}^2$, where $\nu_g = (K - 1)/(1 + \hat{a}_g^2)$, and $\hat{a}_g^2 = \sum_{i=1}^{K-1} (\hat{\lambda}_{g,i} - \hat{\bar{\lambda}}_g)^2 / [(K - 1)\hat{\bar{\lambda}}_g]$. We note that $\hat{\bar{\lambda}}_g = (K - 1)^{-1} \sum_{i=1}^K (\hat{V}_{gii}/\hat{p}_{gi})$ and $\sum_{i=1}^K \lambda_{g,i}^2 = \sum_{i=1}^K \sum_{j=1}^K \{ \hat{V}_{gij}^2 / (\hat{p}_{gi}\hat{p}_{gj}) \}$, where $\hat{V}_{gij}$ is the (i, j) -th element of $\hat{V}_g$. This second-order corrected statistic is better approximated by X^2 than the first-order statistic (3.4.54).

We also note that for the full-sample approach, the application of the second-order statistic (3.4.48) proposed by Rao and Scott (1981) is dependent on the knowledge of $\sum_{i=1}^K \hat{\lambda}_i^2$ or $\hat{V}$ and the public use data file may not contain cluster identification which is needed for calculating $\hat{V}$. On the other hand, the inverse-sampling approach has no such restrictions and the inverse sampling covariance matrix $\hat{V}_g$, given by (3.4.51), can be calculated from the data file consisting of multiple inverse samples. Hence, the second-order corrected statistic (3.4.55) can be implemented from the data file.

Simulation Study

We conducted a limited simulation study to evaluate the performance of the proposed methods for samples of moderately large sizes. For this study we considered a two-stage cluster sampling in which K -category sample of m units is drawn independently from each of r sample clusters. We let $\mathbf{m}_l = (m_{l1}, \dots, m'_{l,K-1})$ represent the vector of counts in the first $K - 1$ categories for the l -th cluster ($l = 1, \dots, r$) and $\mathbf{m} = (m_1, \dots, m_{K-1})'$ represent the corresponding category counts for the whole sample, that is $m_i = \sum_{l=1}^r m_{li}$ ($i = 1, \dots, K - 1$). The total number of observations in the sample is thus $n = mr = \sum_{i=1}^K m_i$, where $m_K = n - \sum_{i=1}^{K-1} m_i$. Further, let

$\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_{K-1})' = \mathbf{m}/n$ be the vector of cell proportions in the sample and define $\mathbf{p} = E(\hat{\mathbf{p}})$, where E denotes the expectation under a specified cluster sampling model. Similarly, let $\mathbf{V}/n$ represent the $(K-1) \times (K-1)$ covariance of $\hat{\mathbf{p}}$, under the model. We compare the performance of the design-based goodness-of-fit test statistics and the proposed inverse-sampling test statistics for testing the null hypothesis $H_0 : \mathbf{p} = \mathbf{p}_0$. The approximate inverse sampling scheme was implemented, i.e., selecting one unit at random from each of the r sample clusters resulting in an inverse sample of size r . Repeat the process to obtain say, g inverse samples each of size r .

The parameters that were controlled for the study are (i) α , the nominal significance level for the tests; (ii) $\mathbf{p}$, the model probability vector; (iii) K , the number of categories; (iv) r , the number of sample clusters; (v) m , the number of units drawn from each sample cluster; (vi) $\bar{\lambda}$, the mean of the eigenvalues λ_i ; (vii) a , the coefficient of variation of the λ_i 's. The degree of clustering is represented by $(\bar{\lambda}, a)$. For this study, we considered $(\bar{\lambda} = 2, a = 0)$, which represents a constant design effect clustering. The nominal significance level used was $\alpha = 0.05$, and the model probability vector under H_0 was $\mathbf{p}_0 = (1/K, \dots, 1/K)'$

Brier (1978) method of generating Dirichlet variates from $K-1$ beta random variables was used. For each of the 1,000 Monte carlo trials, independent Dirichlet vectors, $\mathbf{p}_i$, were generated for $r = (30, 50)$ clusters. For each cluster, a K -category conditional multinomial sample was constructed by referring each m independent $(0, 1)$ uniform variates to the appropriate interval associated with $\mathbf{p}_i$. All the tests methods were applied to these sampled clusters. The results are reported in Table 3.1. We observe that, the inverse-sampling Wald test significance levels track the design

Table 3.1: *Actual significance levels (%) of the design-based and inverse-sampling Wald statistics, X_W^2 and $X_{gc,W}^2$ respectively, $\mathbf{p} = (1/K, 1/K, 1/K)'$; $K = 3$, $\bar{\lambda} = 2$, $a = 0$, $m = 10$, $\alpha = 5\%$.*

Sample clusters	SL(X_W^2)	SL($X_{gc,W}^2$)		
		$g = 500$	$g = 1,000$	$g = 2,000$
50	7.3	6.7	7.7	7.1
30	8.1	7.9	7.8	8.3

g ; number of repeated subsamples

based Wald test significance levels even though the full-sample Wald statistic in this case is poor in tracking the true nominal significance level, 5%. Table 3.2 reports the performance of the first-order and second-order full-sample and inverse-sampling test statistics. We observe that the second-order inverse-sampling statistic, $X_{gc,S}^2$ give

Table 3.2: *Actual significance levels (%) of X_c^2 , X_S^2 , $X_{gc,c}$, $X_{gc,W}^2$, $K = 3$, $\bar{\lambda} = 2$, $a = 0$, $m = 10$, $\alpha = 5\%$, $\mathbf{p} = (1/K, 1/K, 1/K)'$.*

Sample clusters	SL(X_c^2)	SL($X_{gc,c}^2$)			SL(X_S^2)	SL($X_{gc,S}^2$)		
		$g = 500$	1,000	2,000		$g = 500$	1,000	2,000
50	5.7	6.2	6.4	6.9	4.7	4.6	4.4	4.8
30	5.8	8.0	7.5	7.4	4.7	4.9	4.2	5.3

g ; number of repeated subsamples

significance levels close to the nominal level of 5% even for $g(= 500)$. The first-order inverse-sampling significance levels are liberal in the sense of being larger than 5%.

3.4.3 Independence Tests for a Two-way Table

Consider a finite population $U = \{1, \dots, i, \dots, N\}$ of size N . Suppose U is partitioned into $r \times c$ nonoverlapping and exhaustive categories U_{ij} ($i = 1, \dots, r; j = 1, \dots, c$), where r is the number of rows and c is the number of columns. Let p_{ij} be the population proportion in the (i, j) th category, such that $p_{ij} \geq 0$, $\sum_{i=1}^r p_{ij} = 1$ and $\sum_{j=1}^c p_{ij} = 1$. The problem of testing independence between rows and columns in a two-way table is expressed in the hypothesis $H_0 : h_{ij}(\mathbf{p}) = p_{ij} - p_{i+}p_{+j}$, $i = 1, \dots, r-1; j = 1, \dots, c-1$ where $\mathbf{p} = (p_{11}, p_{12}, \dots, p_{rc-1})'$, $p_{i+} = \sum_{j=1}^c p_{ij}$ and $p_{+j} = \sum_{i=1}^r p_{ij}$. We summarize the full-sample methods of testing the hypothesis and then propose inverse-sampling approach as an alternative method to the problem.

Full-sample Independence Tests

Suppose s_0 is a sample of size n drawn from the finite population according to a specified complex design $p_0(s_0)$. A consistent estimator of p_{ij} under $p_0(s_0)$ is given by

$$\hat{p}_{ij} = \frac{\sum_{l \in s_0} y_{ijl} / \pi_l}{\sum_{l \in s_0} 1 / \pi_l}, \quad i = 1, \dots, r; j = 1, \dots, c \quad (3.4.56)$$

where $y_{ijl} = 1$ if $l \in U_{ij}$; 0 otherwise, $\sum_{i=1}^r \hat{p}_{ij} = 1$ and $\sum_{j=1}^c \hat{p}_{ij} = 1$.

The generalized Wald statistic for testing H_0 is given by

$$X_{I(W)}^2 = nh(\hat{\mathbf{p}})' \widehat{\mathbf{V}}_h^{-1} h(\hat{\mathbf{p}}), \quad (3.4.57)$$

where $\widehat{\mathbf{V}}_h/n$ is an estimator of the covariance matrix $\mathbf{V}_h/n$ of $h(\hat{\mathbf{p}})$, $h(\hat{\mathbf{p}})$ is the column vector of $h_{ij}(\hat{\mathbf{p}})$'s. The statistic $X_{I(W)}^2$ is asymptotically distributed as χ_b^2 , under H_0 , where $b = (r-1)(c-1)$ (see Rao and Scott, 1981).

The Pearson chi-squared statistic for testing H_0 is given by

$$\begin{aligned} X_I^2 &= n \sum_{i=1}^r \sum_{j=1}^c \frac{(\hat{p}_{ij} - \hat{p}_{i+} \hat{p}_{+j})^2}{\hat{p}_{i+} \hat{p}_{+j}} \\ &= nh(\hat{\mathbf{p}})' \left(\widehat{\mathbf{P}}_r^{-1} \otimes \widehat{\mathbf{P}}_c^{-1} \right) h(\hat{\mathbf{p}}), \end{aligned} \quad (3.4.58)$$

where $\otimes$ represents Kronecker product, $\widehat{\mathbf{P}}_r$ and $\widehat{\mathbf{P}}_c$ are the values of $\widehat{\mathbf{P}}_r = \text{diag}(\mathbf{p}_r) - \mathbf{p}_r \mathbf{p}_r'$ and $\widehat{\mathbf{P}}_c = \text{diag}(\mathbf{p}_c) - \mathbf{p}_c \mathbf{p}_c'$ respectively, for $\mathbf{p} = \hat{\mathbf{p}}$, $\mathbf{p}_r = (p_{1+}, \dots, p_{r-1,+})'$ and $\mathbf{p}_c = (p_{+1}, \dots, p_{+,c-1})'$. We restate one of the main results of Rao and Scott (1981) in the following theorem.

Theorem 3.8 *Under the null hypothesis $H_0 : h_{ij}(\mathbf{p}) = p_{ij} - p_{i+}p_{+j}$, X_I^2 may be written as $\sum_{i=1}^b \delta_{0i} W_i$ where δ_i 's are the eigenvalues of $\mathbf{D}_h = (\mathbf{P}_r^{-1} \otimes \mathbf{P}_c^{-1}) \mathbf{V}_h$ and δ_{0i} is the value of δ_i under H_0 .*

A first-order correction statistic proposed by Rao and Scott (1981) treats

$$X_{I(c)}^2 = \frac{X_I^2}{\hat{\delta}} \quad (3.4.59)$$

as a chi-squared random variable with b degrees of freedom, where $\hat{\delta} = 1/b \sum_{i=1}^b \hat{\delta}_i$, $\hat{\delta}_i$'s are the eigenvalues of $(\widehat{\mathbf{P}}_r^{-1} \otimes \widehat{\mathbf{P}}_c^{-1}) \widehat{\mathbf{V}}_h$. The statistic $X_{I(c)}^2$ is asymptotically distributed as $Y_I = \sum_{i=1}^b (\delta_{0i}/\bar{\delta}_0) W_i$, whose expectation is b .

We now turn our attention to our propose inverse-sampling independence tests.

Inverse-sampling Independence Tests

Again we focus on cluster sampling. Suppose s_0 is a sample of m clusters drawn from M clusters in a population of N elements according to a one-stage or two-stage cluster sampling design. An inverse sample s^* of size m elements is drawn from s_0 such that

the unconditional probability of s^* , $p(s^*)$, matches simple random sampling either exactly or approximately. Again the exact or approximate inverse sampling schemes for a number of one-stage and two-stage cluster sampling designs are detailed in Subsection 2.2.2 and Subsection 2.2.3 respectively.

The subsampling estimators $\hat{p}_{ijk}^*$, ($k = 1, \dots, g$) of p_{ij} are given by

$$\hat{p}_{ijk}^* = \frac{1}{m} \sum_{l \in s_k^*} y_{ijl}, \quad k = 1, \dots, g \quad (3.4.60)$$

where y_{ijl} is as defined earlier. The inverse-sampling estimator of p_{ij} is given by

$$\hat{p}_{gij} = \frac{1}{g} \sum_{k=1}^g \hat{p}_{ijk}^*. \quad (3.4.61)$$

Let $\hat{\mathbf{p}}_g = (\hat{p}_{g11}, \hat{p}_{g12}, \dots, \hat{p}_{gr,c-1})'$, $\hat{p}_{gi+} = \sum_{j=1}^c \hat{p}_{gij}$ and $\hat{p}_{g+j} = \sum_{i=1}^r \hat{p}_{gij}$.

An inverse-sampling generalized Wald statistic is

$$X_{gIW}^2 = m h(\hat{\mathbf{p}}_g)' \widehat{\mathbf{V}}_{gh}^{-1} h(\hat{\mathbf{p}}_g), \quad (3.4.62)$$

where $h(\hat{\mathbf{p}}_g)$ is the column vector of $h_{ij}(\hat{\mathbf{p}}_g)$'s and

$$\widehat{\mathbf{V}}_{gh} = \widehat{\mathbf{V}} \{h(\hat{\mathbf{p}}_g)\} = \widehat{\mathbf{H}}_g \widehat{\mathbf{V}}_g \widehat{\mathbf{H}}_g', \quad (3.4.63)$$

$\widehat{\mathbf{H}}_g$ is $\partial h(\mathbf{p})/\partial \mathbf{p}'$ evaluated at $\mathbf{p} = \hat{\mathbf{p}}_g$ and

$$\widehat{\mathbf{V}}_g = \widehat{\mathbf{V}}(\hat{\mathbf{p}}_g) = \frac{1}{g} \sum_{k=1}^g \widehat{\mathbf{V}}(\hat{\mathbf{p}}_k^*) - \frac{1}{g} \sum_{k=1}^g (\hat{\mathbf{p}}_k^* - \hat{\mathbf{p}}_g) (\hat{\mathbf{p}}_k^* - \hat{\mathbf{p}}_g)', \quad (3.4.64)$$

$\hat{\mathbf{p}}_k^* = (\hat{p}_{k11}^*, \hat{p}_{k12}^*, \dots, \hat{p}_{kr,c-1}^*)'$, and $\widehat{\mathbf{V}}(\hat{\mathbf{p}}_k^*) = 1/m \{\text{diag}(\hat{\mathbf{p}}_k^*) - \hat{\mathbf{p}}_k^* \hat{\mathbf{p}}_k^{*'}\}$. The following theorem provides an asymptotic distribution of the inverse-sampling statistic under the null hypothesis.

Theorem 3.9 *Under the null hypothesis $H_0 : h_{ij}(\mathbf{p}) = p_{ij} - p_{i+}p_{+j}$ and provided mild regularity conditions are satisfied, then the inverse-sampling generalized Wald statistic X_{gIW}^2 converges to a χ_b^2 as $g \rightarrow \infty$ and then $m \rightarrow \infty$.*

Proof

Assume that $\mathbf{H} = \frac{\partial \mathbf{h}}{\partial \mathbf{p}'}$ is continuous in the neighbourhood of the true parameter, $\mathbf{p}$ and that $\mathbf{H}$ has rank b . Then we may write using Taylors series expansion

$$m^{\frac{1}{2}} \{h(\hat{\mathbf{p}}_g) - h(\mathbf{p})\} \approx \mathbf{H}(\mathbf{p})m^{\frac{1}{2}} (\hat{\mathbf{p}}_g - \mathbf{p}). \quad (3.4.65)$$

Apply the results of Theorem 2.7 to the components of $\hat{\mathbf{p}}_g$, then it easily follows that $m^{\frac{1}{2}} (\hat{\mathbf{p}}_g - \mathbf{p})$ is asymptotically distributed as normal with mean $\mathbf{0}$ and covariance $\mathbf{V}$. Hence by (3.4.65), $m^{\frac{1}{2}} \{h(\hat{\mathbf{p}}_g) - h(\mathbf{p})\}$ is asymptotically normally distributed with zero mean and covariance $\mathbf{H}(\mathbf{p})\mathbf{V}\mathbf{H}(\mathbf{p})'$. The rest of the proof easily follows from standard results. ■

An inverse-sampling Pearson statistic is given by

$$\begin{aligned} X_{gI}^2 &= m \sum_{i=1}^r \sum_{j=1}^c \frac{(\hat{p}_{gij} - \hat{p}_{gi+}\hat{p}_{g+j})^2}{\hat{p}_{gi+}\hat{p}_{g+j}} \\ &= m h(\hat{\mathbf{p}}_g)' \left(\widehat{\mathbf{P}}_{gr}^{-1} \otimes \widehat{\mathbf{P}}_{gc}^{-1} \right) h(\hat{\mathbf{p}}_g), \end{aligned} \quad (3.4.66)$$

where $\otimes$ is the direct matrix product, $\hat{\mathbf{p}}_g$, $h(\hat{\mathbf{p}}_g)$ are as defined before, $\widehat{\mathbf{P}}_{gr}$ and $\widehat{\mathbf{P}}_{gc}$ are the values of $\mathbf{P}_r = \text{diag}(\mathbf{p}_r) - \mathbf{p}_r \mathbf{p}_r'$ and $\mathbf{P}_c = \text{diag}(\mathbf{p}_c) - \mathbf{p}_c \mathbf{p}_c'$ respectively for $\mathbf{p} = \hat{\mathbf{p}}_g$, $\mathbf{p}_r = (p_{r+}, \dots, p_{r-1,+})'$ and $\mathbf{p}_c = (p_{+c}, \dots, p_{+,c-1})'$.

The theorem below gives the asymptotic distribution of X_{gI}^2 .

Theorem 3.10 Under the null hypothesis $H_0 : h_{ij}(\mathbf{p}) = p_{ij} - p_{i+}p_{+j}$, $i = 1, \dots, r - 1$, $j = 1, \dots, c - 1$, X_{gI}^2 converges in distribution to $Y = \sum_{i=1}^b \delta_{0i} W_i$ as $g \rightarrow \infty$ then $m \rightarrow \infty$, where δ_i 's are the eigenvalues of $D_h = (P_r^{-1} \otimes P_c^{-1}) (H V H') = (P_r^{-1} \otimes P_c^{-1}) V_h$ and δ_{0i} 's is the value of δ_i under H_0 , W_i 's are independent χ_1^2 .

Proof

Under H_0 , we note that

$$X_{gI}^2 = \left[m^{\frac{1}{2}} \widehat{V}_h^{-\frac{1}{2}} \{h(\hat{\mathbf{p}}_g) - h(\mathbf{p}_0)\} \right]' \widehat{V}_h^{\frac{1}{2}} (P_{r0}^{-1} \otimes P_{c0}^{-1}) \widehat{V}_h^{\frac{1}{2}} \left[m^{\frac{1}{2}} \widehat{V}_h^{-\frac{1}{2}} \{h(\hat{\mathbf{p}}_g) - h(\mathbf{p}_0)\} \right] \quad (3.4.67)$$

The proof follows similar arguments as in the proof of Theorem 3.9. In addition using the results of Theorem 2.8, it follows that $m^{\frac{1}{2}} \{h(\hat{\mathbf{p}}_g) - h(\mathbf{p}_0)\}$ is asymptotically normally distributed with zero mean and covariance V_h . The rest of the proof easily follows from standard results. ■

An inverse-sampling first-order correction statistic we propose is

$$X_{gI(c)}^2 = \frac{X_{gI}^2}{\hat{\delta}_g}, \quad (3.4.68)$$

where $\hat{\delta}_g = 1/b \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \hat{\delta}_{gij}$ and $\hat{\delta}_{gij} = \hat{V}_{h_{gij}} / (b \hat{p}_{gi+} \hat{p}_{g+j})$. We note that $X_{gI(c)}^2$ is distributed asymptotically as $Y_1 = \sum_{i=1}^b \delta_{0i} W_i$. The expectation of Y_1 is b which is the same expectation as χ_b^2 , the variance of Y_1 is $V(Y_1) = 2b + 2 \sum_{i=1}^b (\delta_{0i} - \bar{\delta}_0)^2 / \bar{\delta}_0^2$ which is larger than $V(\chi_b^2) = 2b$ unless all the δ_{0i} 's are equal. Thus treating $X_{gI(c)}$ as χ_{K-1}^2 under H_0 will tend to give liberal significance test estimates.

A second-order inverse-sampling correction statistic we propose is given by

$$X_{gI(S)} = \frac{X_{gI(c)}^2}{1 + \hat{a}_{gI}^2}, \quad (3.4.69)$$

which is treated as χ_ν^2 , where $\nu = b/(1 + \hat{a}_{gI}^2)$, $\hat{a}_{gI}^2 = \sum_{i=1}^b (\hat{\delta}_{gi} - \hat{\delta}_g)^2 / (b\hat{\delta}_g^2)$. This second-order corrected statistic is better approximated by χ^2 than the first-order statistic.

3.5 Quasi-Score Tests

For a finite population of size N , let $(\mathbf{x}_i, y_i)$ be all the units measured on a vector of explanatory variables, $\mathbf{X}$, and a response variable, Y . We assume that for a given value of $\mathbf{x}$, Y is generated by some random process with mean

$$E(Y_i) = \mu_i = \mu(\mathbf{x}_i, \boldsymbol{\beta}), \quad (3.5.70)$$

and we suppose that we have in mind some working model for the variance, say

$$Var(Y_i) = V_{0i} = V_0(\mu_i), \quad (3.5.71)$$

for $i = 1, \dots, N$. The model (3.5.70) for the mean is assumed to be valid, but the working variance (3.5.71) may only be rough approximation at best.

We do not observe values for all the population units but only for those in a sample drawn from the finite population according to some well defined sampling scheme. We are interested in testing the hypothesis $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}$ using the sample data, where $\boldsymbol{\beta}$ is partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$ with $\boldsymbol{\beta}_2$ a $q \times 1$ vector, and $\boldsymbol{\beta}_1$ a $r \times 1$ vector.

Using the finite population estimating equations

$$U(\boldsymbol{\beta}) = \sum_{i=1}^N \mathbf{u}_i(\boldsymbol{\beta}) = \mathbf{0}, \quad (3.5.72)$$

where $u_i(\beta)$ has a k -th component $u_{ik} = (\partial\mu_i/\partial\beta_k)(y_i - \mu_i)/V_{0i}$, we could obtain a consistent estimator of β .

In reality, we do not know the values for the whole finite population but only for those units in a sample drawn from the population. We suppose only that the sample design provides consistent, asymptotically normal estimators of population totals, and associated standard errors. Then since $U(\beta)$ is a vector of population totals for fixed β , we can produce an estimator of $U(\beta)$ as

$$\widehat{U}(\beta) = \sum_{i \in s_0} w_i u_i(\beta), \quad (3.5.73)$$

where w_i 's are the survey weights, here we focus on $w_i = 1/\pi_i$, and π_i is the inclusion probability of the i -th unit. A sample estimator, $\hat{\beta}$ is obtained by solving $\widehat{U}(\hat{\beta}) = 0$.

Under suitable regularity conditions (Binder, 1983), $\hat{\beta}$ is shown to be asymptotically normal with mean β , and covariance matrix $Cov(\hat{\beta})$ consistently estimated by

$$\widehat{V}(\hat{\beta}) = \left\{ \frac{\partial \widehat{U}(\beta)}{\partial \beta'} \bigg|_{\beta=\hat{\beta}} \right\}^{-1} \widehat{V} \{ \widehat{U}(\hat{\beta}) \} \left\{ \frac{\partial \widehat{U}(\beta)}{\partial \beta'} \bigg|_{\beta=\hat{\beta}} \right\}^{-1}, \quad (3.5.74)$$

where $\widehat{V} \{ \widehat{U}(\hat{\beta}) \}$ is the estimated covariance matrix of $\widehat{U}(\beta)$ under the specified design evaluated at $\beta = \hat{\beta}$. Note that $\widehat{V} \{ \widehat{U}(\hat{\beta}) \}$ is obtained from the standard survey variance estimator for a total since $\widehat{U}(\beta)$, given by (3.5.73), is the estimator of the total $U(\beta)$.

For the problem of testing the null hypothesis, $H_0 : \beta_2 = \beta_{20}$, one approach is to base the test on the corresponding Wald statistic

$$X_W^2 = (\hat{\beta}_2 - \beta_{20})' \widehat{V}(\hat{\beta}_2)^{-1} (\hat{\beta}_2 - \beta_{20}) \quad (3.5.75)$$

which is asymptotically a χ_q^2 random variable under H_0 .

An alternative test with more desirable properties (see Rao, Scott and Skinner, 1998) is the score test. We summarize their result. Let $\tilde{\beta} = (\tilde{\beta}_1', \beta_{20}')'$ be the solution of $\widehat{U}_1(\tilde{\beta}) = 0$ where $\widehat{U} = (\widehat{U}_1', \widehat{U}_2')'$ is partitioned in the same way as β . The analogue of the score test, which they referred to as quasi-score test, is based on the statistic

$$X_S^2 = \widetilde{U}_2' \widetilde{V}_{2S}^{-1} \widetilde{U}_2, \quad (3.5.76)$$

where $\widetilde{U}_2 = \widehat{U}_2(\tilde{\beta})$ and $\widetilde{V}_{2S}$ is a design consistent estimator of $Cov(\widetilde{U}_2)$. Rao, Scott and Skinner (1998) showed (heuristically) that $\widetilde{U}_2$ is asymptotically normal with mean 0 and covariance matrix $Cov(\widetilde{U}_2)$ under H_0 . Thus X_S^2 is asymptotically a χ_q^2 random variable under H_0 .

Estimation of $Cov(\widetilde{U}_2)$ as shown in Rao, Scott and Skinner (1998), may be done using Taylor linearization estimation approach or resampling methods such as the jackknife or the balanced repeated replication.

The preceding methods for estimating the covariance $Cov(\widetilde{U}_2)$, require the use of sampling weights, cluster identifiers as well as specialized computer softwares for their implementation. For public use microdata file, cluster identifiers are not usually available due to the need to protect the confidentiality of respondents. We propose alternative methods based on inverse sampling to overcome some of these limitations.

Inverse-sampling hypothesis tests

We propose two inverse-sampling methods for testing the hypothesis, $H_0 : \beta = \beta_{20}$, namely, (i) inverse-sampling Wald test; (ii) inverse-sampling quasi-score test.

(i) Inverse-sampling Wald Test

The combined estimating equation estimator, $\tilde{\beta}_{gc} = (\tilde{\beta}'_{gc1}, \beta'_{20})'$ of $\beta = (\beta'_1, \beta'_2)'$ under H_0 , is the solution to $\widehat{U}_{gc}(\beta) = 0$, where the combined estimating function, $\widehat{U}_{gc}(\beta)$ is given by (3.2.9). Assuming that $(\hat{\beta}_{gc2} - \beta_2)$ is normally distributed with mean 0 and covariance matrix Σ_2 , then the inverse-sampling Wald statistic

$$X^2_{gc,W} = (\hat{\beta}_{gc2} - \beta_2)' \widehat{V}(\hat{\beta}_{gc2})^{-1} (\hat{\beta}_{gc2} - \beta_2) \quad (3.5.77)$$

under H_0 , is asymptotically a χ^2_q random variable with q degrees of freedom and $\widehat{V}(\hat{\beta}_{gc2})$ is the inverse-sampling covariance estimator of $\hat{\beta}_{gc2}$ given by (3.2.14) with $\hat{\theta}_{gc}$ replaced by $\hat{\beta}_{gc2}$. The statistic (3.5.77) has the usual problems associated with the Wald test. For example, it is not invariant to nonlinear transformations of the parameters β and often has poor small sample behaviour. The covariance estimator, $\widehat{V}(\hat{\beta}_{gc2})$ may also be unstable when the dimension of β_2 is large.

An alternative statistic with more desirable properties is the inverse-sampling quasi-score test.

(ii) Inverse-sampling quasi-score test

Let $\widehat{U}_{gc}(\beta) = (\widehat{U}'_{gc1}, \widehat{U}'_{gc2})'$ be a partition corresponding to $\beta = (\beta'_1, \beta'_2)'$ and $\tilde{\beta}_{gc} = (\tilde{\beta}'_{gc1}, \beta'_{20})'$ the solution to $\widehat{U}_{gc1}(\tilde{\beta}_{gc}) = 0$, where $\widehat{U}_{gc}(\beta)$ is the combined estimating function given in (3.2.9). Suppose $\widetilde{U}_{gc2} = \widehat{U}_{gc2}(\tilde{\beta}_{gc})$ is asymptotically normal with mean 0 and covariance matrix Σ_{2U} under H_0 . Then the inverse-sampling quasi-score statistic

$$X^2_{gc,S} = \widetilde{U}'_{gc2} \widetilde{V}_{gc2U}^{-1} \widetilde{U}_{gc2}, \quad (3.5.78)$$

under H_0 , is asymptotically a χ^2_q random variable with q degrees of freedom, where

the inverse-sampling covariance estimator, $\widetilde{V}_{gc2U}^{-1}$ is given by

$$\begin{aligned}\widetilde{V}_{gc2U} &= \widehat{V}(\widetilde{U}_{gc2}) \\ &= \frac{1}{g} \sum_{j=1}^g \widehat{V}_{jU}^* - \frac{1}{g} \sum_{j=1}^g \left\{ \widehat{U}_{j2}^*(\widetilde{\beta}_{gc}) - \widehat{U}_{gc2}(\widetilde{\beta}_{gc}) \right\} \left\{ \widehat{U}_{j2}^*(\widetilde{\beta}_{gc}) - \widehat{U}_{gc2}(\widetilde{\beta}_{gc}) \right\}',\end{aligned}\tag{3.5.79}$$

where

$$\widehat{V}_{jU}^* = \frac{N^2}{m} \left(1 - \frac{m}{N}\right) \frac{1}{m-1} \sum_{i \in s_j^*} (\hat{z}_i - \hat{\bar{z}})(\hat{z}_i - \hat{\bar{z}})',$$

$\hat{z}_i$ is $z_i = u_{2i}(\beta) - Au_{1i}(\beta)$ evaluated at $\beta = \widetilde{\beta}_{gc}$, u_{1i} and u_{2i} are the components of the partition of u_i corresponding to $\beta = (\beta'_1, \beta'_2)'$, u_i is the i -th estimand of the combined estimating function $\widehat{U}_{gc}(\beta)$, I_{11} and I_{12} are the first column components of the matrix $I(\beta) = -E(\partial \widehat{U}_{gc}(\beta) / \partial \beta')$, and $\widehat{U}_{j2}^*(\widetilde{\beta}_{gc})$ is $\widetilde{U}_{j2}^*(\beta) = \sum_{i \in s_j^*} \hat{z}_i / m$ evaluated at $\beta = \widetilde{\beta}_{gc}$.

Part II

Analyses of Cluster-Correlated Biological Data Using Inverse Sampling

Chapter 4

Analysis of Cluster-Correlated Data with Nonignorable Cluster Sizes

4.1 Introduction

In the previous chapters we presented some theory and application of inverse sampling in the context of complex survey sampling. In Chapter 3 we studied a combined estimating equation approach to inverse sampling and demonstrated that it leads to asymptotically valid inferences for nonlinear parameters such as linear regression and logistic regression parameters in the context of finite population. The application of this method extends beyond finite populations. In this chapter we show how to use inverse-sampling combined estimating equation and mean estimating equation methods to analyze cluster-correlated biological data in the context of infinite population

or model-based framework.

It is well known that methods for analyzing cluster-correlated biological data that treat positively correlated outcomes as independent tend to produce standard errors that are too small and hypothesis tests that have an inflated Type I error rate. Many statistical techniques have been developed to handle cluster-correlated outcomes. A widely used method, due to Liang & Zeger (1986), uses generalized estimating equations under “working” independence or correlation within clusters. This method implicitly assumes the ignorability of cluster sizes. However, in a number of applications this may not be true. For example, consider a study of the effects of toxicants on the pups of dams and interest is to estimate the association between the explanatory variables (toxicants) and the birth defect status of a pup. Dams that are particularly susceptible to the effects of toxicants are more likely to experience foetal deaths. As well their pups are also more likely to have birth defects. Thus the disease status may be negatively related to the number of pups. Here the cluster size (number of pups) is nonignorable. Another example is the study of periodontal disease. The disease status of teeth from the same person will be correlated, but the number of teeth may be negatively related to disease status because persons who are more susceptible to the disease may already have lost some teeth from the disease. Again in this study, the cluster size (the number of teeth per person) is nonignorable.

When cluster sizes are nonignorable, the Liang-Zeger method may lead to asymptotically invalid inferences. Hoffman, Sen & Weinberg (2001) proposed a simple but computationally intensive method, based on a large number of within-cluster resamples (inverse samples) and associated separate estimating equations that leads to

asymptotically valid inferences under both ignorable and nonignorable cluster sizes. In this method, one observation is chosen at random from each cluster to produce a subsample of independent observations which may be analyzed by standard software. The subsampling procedure is repeated a large number of times, say g , to produce g separate estimating equations with parameter estimators $\hat{\beta}_1^*, \dots, \hat{\beta}_j^*, \dots, \hat{\beta}_g^*$ and associated variance estimators, $\widehat{V}_1^*, \dots, \widehat{V}_j^*, \dots, \widehat{V}_g^*$. The final estimator of the parameter β is then given by $\hat{\beta}_{gS} = g^{-1} \sum_{j=1}^g \hat{\beta}_j^*$ with the variance estimator

$$\widehat{V}_{gS} = \frac{1}{g} \sum_{j=1}^g \widehat{V}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\beta}_j^* - \hat{\beta}_{gS}) (\hat{\beta}_j^* - \hat{\beta}_{gS})'. \quad (4.1.1)$$

Hoffman et al. (2001) demonstrated through a simulation study that their method yields valid parameter estimates and associated standard errors unlike the Liang-Zeger method. They also provided asymptotic justification of their method. We propose two alternative methods: (i) combined estimating equation, (ii) mean estimating equation. In the combined estimating equation method, an observation is chosen at random from each cluster to obtain a subsample of independent observations. The procedure is repeated many times and then the separate estimating equations are combined before solving for the parameters of the model. The mean estimating equation method is a simple method that avoids resampling, and it is motivated by the combined estimating equation method. Standard software such as Splus may easily be modified to implement the mean estimating method.

The rest of the chapter is organized as follows. Section 4.2 gives a brief discussion of the Liang-Zeger method based on working independence within clusters. A mean model, a nested error linear regression model and a logistic linear regression model

with a binary cluster-specific covariate are used to study the bias of Liang-Zeger estimators under nonignorable cluster sizes. The separate estimating equations method of Hoffman et al. (2001), is summarized in Section 4.3. The combined estimating equation method and the mean estimating equation method are introduced in Section 4.4 and Section 4.5 respectively. The efficiency of the mean estimating equation method relative to the Liang-Zeger method is studied in Section 4.6 for the three examples under both ignorable and nonignorable cluster sizes. Section 4.7 gives Wald tests for ignorability of cluster sizes by comparing the mean estimating equation estimator with the corresponding Liang-Zeger estimator. Limited simulation results are presented in Section 4.8 to assess the performance of the proposed methods.

4.2 Method of Liang-Zeger

Consider the cluster-correlated sample $s = \{(\mathbf{x}_{ij}, y_{ij}) : j = 1, \dots, n_i; i = 1, \dots, K\}$, where $\mathbf{x}_{ij}$ is a $p \times 1$ vector of fixed covariates attached to the j -th individual in the i -th cluster, y_{ij} is the corresponding outcome variable which may depend on the cluster size n_i , and K is the number of clusters. Assume that the marginal model of interest is defined by

$$E(y_{ij}) = \mu(\mathbf{x}_{ij}'\boldsymbol{\beta}) = \mu(\eta_{ij}), \quad (4.2.2)$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters.

Under model (4.2.2), the regression parameter $\boldsymbol{\beta}$ may be estimated using working

independence estimating equations

$$U_I(\beta) = \frac{1}{N} \sum_{i=1}^K \sum_{j=1}^{n_i} \frac{(y_{ij} - \mu_{ij}) \mathbf{x}_{ij}}{a_{ij}} \frac{\partial \mu_{ij}}{\partial \eta_{ij}} = \frac{1}{N} \sum_{i=1}^K \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\beta) = \mathbf{0}, \quad (4.2.3)$$

where $a_{ij} = \text{var}(y_{ij})$ and $N = \sum_{i=1}^K n_i$. Equation (4.2.3) is based on the working assumption that the observations within a cluster are independent of one another and the cluster sizes are ignorable. In general, (4.2.3) is solved iteratively to obtain the solution $\hat{\beta}_I$; in particular, the Newton-Raphson method is used. The r -th step of the algorithm is given by

$$\hat{\beta}^{(r)} = \hat{\beta}^{(r-1)} + \{J_I(\hat{\beta}^{(r-1)})\}^{-1} U_I(\hat{\beta}^{(r-1)}), \quad (4.2.4)$$

where $\hat{\beta}^{(r-1)}$ is the value of β obtained at the $(r-1)$ -th iteration, and $U_I(\hat{\beta}^{(r-1)})$ and $J_I(\hat{\beta}^{(r-1)})$ are the values of $U_I(\beta)$ and

$$J_I(\beta) = -\frac{\partial U_I(\beta)}{\partial \beta'} = \frac{1}{N} \sum_{i=1}^K \sum_{j=1}^{n_i} \frac{\mathbf{x}_{ij} \mathbf{x}_{ij}'}{a_{ij}} \left(\frac{\partial \mu_{ij}}{\partial \eta_{ij}} \right)^2$$

evaluated at $\beta = \hat{\beta}^{(r-1)}$. Iterating the algorithm to convergence produces the estimator $\hat{\beta}_I$ as well as $J_I(\hat{\beta}_I)$, the observed information matrix divided by the total number of observations, N .

Under cluster size ignorability, $\hat{\beta}_I$ is consistent for β as $K \rightarrow \infty$. Further, a consistent estimator of the covariance matrix of $\hat{\beta}_I$ is given by

$$\widehat{V}_I(\hat{\beta}_I) = \{J_I(\hat{\beta}_I)\}^{-1} \left[\frac{1}{N^2} \sum_{i=1}^K \left\{ \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\hat{\beta}_I) \right\} \left\{ \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\hat{\beta}_I) \right\}' \right] \{J_I(\hat{\beta}_I)\}^{-1}.$$

The estimator $\hat{\beta}_I$ may not be consistent if the cluster sizes are nonignorable, as shown below for three examples. Nonignorable cluster sizes, n_i , imply the violation of the conditional independence property $E(y_{ij}|n_i) = E(y_{ij})$.

Example 4.1 Mean model

Consider a mean model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}; j = 1, \dots, n_i; i = 1, \dots, K, \quad (4.2.5)$$

where $\mu = E(y_{ij})$, y_{ij} is the response for the j -th individual in the i -th cluster, α_i is the i -th cluster effect and $n_i = h(\alpha_i)$ is the i -th cluster size. We further assume that the random effects, α_i , and the errors, ϵ_{ij} , are independent and identically distributed with mean 0 and variances σ_α^2 and σ_ϵ^2 respectively. The Liang-Zeger independence estimating equations estimator of μ is given by

$$\hat{\mu}_I = \left(\sum_{i=1}^K n_i \right)^{-1} \sum_{i=1}^K n_i \bar{y}_i, \quad (4.2.6)$$

where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$. We note that $\hat{\mu}_I \rightarrow_p \tilde{\mu} = \mu + \bar{n}^{-1} \text{cov}(n_i, \bar{y}_i)$, where $\bar{n} = E(n_i)$. This means that under the above set-up of nonignorable cluster sizes, $\hat{\mu}_I$ is inconsistent unless the n_i 's and y_{ij} 's are uncorrelated.

Example 4.2 Nested error linear regression model

Consider a nested error linear regression model

$$y_{ij} = \mathbf{x}_i' \boldsymbol{\beta} + \alpha_i + \epsilon_{ij}; j = 1, \dots, n_i; i = 1, \dots, K, \quad (4.2.7)$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters, $\mathbf{x}_i$ is a $p \times 1$ vector of fixed cluster-specific covariates, y_{ij} is the response for the j -th individual in the i -th cluster, α_i is the i -th cluster effect and n_i is as defined in Example 4.1. The model (4.2.7) leads to the marginal model $E(y_{ij} | \mathbf{x}_i) = \mu_i = \mathbf{x}_i' \boldsymbol{\beta}$.

The Liang-Zeger independence estimating equations estimator of β for the marginal model $E(y_{ij}|\mathbf{x}_i) = \mathbf{x}_i'\beta$ is given by

$$\hat{\beta}_I = \left(\sum_{i=1}^K n_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^K \sum_{j=1}^{n_i} \mathbf{x}_i y_{ij}.$$

As $K \rightarrow \infty$, $E(\hat{\beta}_I)$ converges to

$$\tilde{\beta} = \beta + \frac{E\{\alpha h(\alpha)\}}{E\{h(\alpha)\}} \left(\sum_{i=1}^K \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^K \mathbf{x}_i. \quad (4.2.8)$$

In general, the last term of (4.2.8) is not a zero vector. As a result, $\hat{\beta}_I$ is asymptotically biased for β . However, for models with an intercept term, β_0 , we have

$$\left(\sum_{i=1}^K \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^K \mathbf{x}_i = (1, 0, \dots, 0)'. \quad (4.2.9)$$

The result (4.2.9) follows by letting $\mathbf{x}_i' = (1, \mathbf{z}_i')$ and noting that

$$\sum_{i=1}^K \mathbf{x}_i \mathbf{x}_i' = K \begin{pmatrix} 1 & b' \\ b & C \end{pmatrix}, \quad \sum_{i=1}^K \mathbf{x}_i = K \begin{pmatrix} 1 \\ b \end{pmatrix}$$

and

$$\begin{pmatrix} 1 & b' \\ b & C \end{pmatrix}^{-1} \begin{pmatrix} 1 & b' \\ b & C \end{pmatrix} = \begin{pmatrix} 1 & 0' \\ 0 & I \end{pmatrix},$$

where $b = K^{-1} \sum_{i=1}^K \mathbf{z}_i$, I is $(K-1) \times (K-1)$ identity matrix and $C = K^{-1} \sum_{i=1}^K \mathbf{z}_i \mathbf{z}_i'$.

Combining (4.2.8) with (4.2.9), it now follows that only the intercept estimator, $\hat{\beta}_{0I}$, is asymptotically biased under the above set-up of nonignorable cluster sizes for marginal linear regression models with an intercept term. In practice, investigators are typically interested in the estimators of covariate effects, i.e., the slope parameter estimators, and not the intercept estimator. However, the bias in the intercept term affects estimator of the mean responses, $E(y_{ij})$, at selected values of the covariates $\mathbf{x}$.

Example 4.3 *Logistic linear regression model*

Consider the logistic linear regression model

$$\text{logit}\{\text{pr}(y_{ij} = 1|z_i)\} = \beta_0 + \beta_1 z_i, \quad j = 1, \dots, n_i; \quad i = 1, \dots, K, \quad (4.2.10)$$

with cluster-specific covariate z_i taking values 0 or 1, where $\text{logit}(a) = \log\{a(1 - a)^{-1}\}$, $0 < a < 1$. We can write $\beta_0 = \text{logit}(\pi_0)$ and $\beta_1 = \text{logit}(\pi_1) - \text{logit}(\pi_0)$, where $\pi_l = \text{pr}(y = 1|z = l)$, $l = 0, 1$. Estimation of β_0 and β_1 is thus equivalent to estimating π_0 and π_1 . The independence estimating equations estimator of π_l is

$$\hat{\pi}_{lI} = \left(\sum_{S_l} n_i \right)^{-1} \left(\sum_{S_l} n_i \bar{y}_i \right), \quad l = 0, 1,$$

where $\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ and S_l denotes the set of sample clusters with $z_i = l$ for $l = 0, 1$. Hence, the independence estimating equations estimators of β_0 and β_1 are given by $\hat{\beta}_{0I} = \text{logit}(\hat{\pi}_{0I})$ and $\hat{\beta}_{1I} = \text{logit}(\hat{\pi}_{1I}) - \text{logit}(\hat{\pi}_{0I})$. Under nonignorable cluster sizes, $\hat{\pi}_{lI} \rightarrow \tilde{\pi}_l = \pi_l + \bar{n}_l^{-1} \text{cov}_l(n_i, \bar{y}_i)$ with $\bar{n}_l = E(n_i|z_i = l)$, where cov_l denotes covariance given $z_i = l$. It follows that $\tilde{\pi}_l$ is not equal to π_l unless the n_i 's and y_{ij} 's are uncorrelated given $z_i = l$.

We note that $\text{logit}(\tilde{\pi}_l) = \text{logit}(\pi_l) + a_l$, where

$$a_l = \log \left[\frac{1 + (\pi_l \bar{n}_l)^{-1} \text{cov}_l(n_i, \bar{y}_i)}{1 - \{(1 - \pi_l) \bar{n}_l\}^{-1} \text{cov}_l(n_i, \bar{y}_i)} \right],$$

so that $\tilde{\beta}_0 = \beta_0 + a_0$ and $\tilde{\beta}_1 = \beta_1 + (a_1 - a_0)$. Thus $\hat{\beta}_{1I}$ will be approximately unbiased for the slope β_1 if $|a_1 - a_0|$ is small. It also follows from $\tilde{\pi}_l = \pi_l + \bar{n}_l^{-1} \text{cov}_l(n_i, \bar{y}_i)$ that $\tilde{\pi}_l < \pi_l$ and $a_0 < 0$ if $\text{cov}_0(n_i, \bar{y}_i) < 0$. Hence, $\hat{\beta}_{0I}$ underestimates β_0 in this case. The bias of the slope estimator $\hat{\beta}_{1I}$ may not be small if $\text{cov}_l(n_i, \bar{y}_i)$, $l = 0, 1$

have opposite signs or if a_1 differs significantly from a_0 . For the simulation study in Section 4.8 based on model (4.8.37), we have $a_1 = -0.382$ and $a_0 = -0.397$ so that $\hat{\beta}_{1I}$ is approximately unbiased for β_1 . On the other hand, $\beta_{0I} - \beta_0 = a_0 = -0.397$ so that $\hat{\beta}_{0I}$ underestimates β_0 .

To illustrate the asymptotic bias of $\hat{\beta}_{0I}$ and $\hat{\beta}_{1I}$, suppose that $E(y_{ij}|b_i, z_i = 0) = b_i$, $E(y_{ij}|b_i, z_i = 1) = 0.8b_i + 0.3$ and $n_i = n(b_i) = 1.5b_i^{-1}$, where b_i takes the values 0.25 or 0.75 with probability $\frac{1}{2}$. Then $\pi_0 = 0.5$ and $\pi_1 = 0.7$ so that $\beta_0 = 0$ and $\beta_1 = 0.847$. In addition, $\bar{n}_i = \bar{n} = 4$, $\text{cov}_0(n_i, \bar{y}_i) = -0.5$ and $\text{cov}_1(n_i, \bar{y}_i) = -0.4$. It follows that $\pi_{0I} = 0.375$ so that $\hat{\beta}_{0I} \rightarrow -0.51$, and $\pi_{1I} = 0.6$ so that $\hat{\beta}_{1I} \rightarrow 0.916$. This example illustrates that both $\hat{\beta}_{0I}$ and $\hat{\beta}_{1I}$ can be asymptotically biased, although $\hat{\beta}_{1I}$ has small bias here. It is easy to construct examples where the slope estimator $\hat{\beta}_{1I}$ has large bias or both estimators $\hat{\beta}_{0I}$ and $\hat{\beta}_{1I}$ have large biases. Suppose that $E(y_{ij}|b_i, z_i = 0) = b_i$, $E(y_{ij}|b_i, z_i = 1) = 1 - b_i$ and $n_i = n(b_i) = 10b_i + 1$, where b_i takes the value 0.1 with probability 0.9 or the value 0.9 with probability 0.1. Then $\pi_0 = 0.18$ and $\pi_1 = 0.82$ so that $\beta_0 = -1.516$ and $\beta_1 = 3.032$. In addition, $\bar{n}_i = \bar{n} = 2.8$, $\text{cov}_0(n_i, \bar{y}_i) = 0.576$ and $\text{cov}_1(n_i, \bar{y}_i) = -0.576$. It follows that $\pi_{0I} = 0.386$ so that $\hat{\beta}_{0I} \rightarrow -0.464$ and $\pi_{1I} = 0.614$ so that $\hat{\beta}_{1I} \rightarrow 0.928$.

4.3 Separate Estimating Equations

In this section, we summarize the separate estimating equations method proposed by Hoffman et al. (2001). We simply select one unit at random from each cluster to obtain an independent subsample of size K and then repeat this process a large

number of times, say g . The resulting subsamples $s_j^* = \{(x_{ij}^*, y_{ij}^*), i = 1, \dots, K\}$; $j = 1, \dots, g$ are conditionally independent given s . Resampling estimators, $\hat{\beta}_j^*$, are obtained by solving g estimating equations separately:

$$U_j^*(\beta) = \frac{1}{K} \sum_{i=1}^K \left(\frac{\partial \mu_{ij}^*}{\partial \beta} \right)' \frac{1}{a_{ij}^*} (y_{ij}^* - \mu_{ij}^*) = \frac{1}{K} \sum_{i=1}^K u_{ij}^*(\beta) = 0; j = 1, \dots, g, \quad (4.3.11)$$

where $\mu_{ij}^* = E(y_{ij}^*)$ and $a_{ij}^* = \text{var}(y_{ij}^*)$. In general, we require g iterative solutions to get $\hat{\beta}_1^*, \dots, \hat{\beta}_g^*$ from (4.3.11). The separate estimating equations estimator of β is given by

$$\hat{\beta}_{gS} = \frac{1}{g} \sum_{j=1}^g \hat{\beta}_j^*. \quad (4.3.12)$$

Note that the individual estimators, $\hat{\beta}_j^*$ in (4.3.12) may be unstable if $K(> p)$ is small because $\hat{\beta}_j^*$ is based only on K independent observations. The inverse-sampling covariance estimator of $\hat{\beta}_{gS}$, proposed by Hoffman et al. (2001), is given by

$$\widehat{V}_{gS} = \frac{1}{g} \sum_{j=1}^g \widehat{V}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\beta}_j^* - \hat{\beta}_{gS})(\hat{\beta}_j^* - \hat{\beta}_{gS})', \quad (4.3.13)$$

where $\widehat{V}_j^*$ is the “sandwich” variance estimator of $\hat{\beta}_j^*$:

$$\widehat{V}_j^* = \{J_j^*(\hat{\beta}_j^*)\}^{-1} \left\{ \frac{1}{K^2} \sum_{i=1}^K u_{ij}^*(\hat{\beta}_j^*) u_{ij}^*(\hat{\beta}_j^*)' \right\} \{J_j^*(\hat{\beta}_j^*)\}^{-1}, \quad (4.3.14)$$

where $J_j^*(\hat{\beta}_j^*)$ is $J_j^*(\beta) = -\partial U_j^*(\beta) / \partial \beta'$ evaluated at $\hat{\beta}_j^*$. The variance estimator $\widehat{V}_{gS}$ may not be positive definite if K is small, even for large g . As a result, the variance estimators of the components of $\hat{\beta}_{gS}$ can take negative values quite often if K is small; see Subsection 4.8.4.

Assuming large g , Hoffman et al. (2001) proved that $K^{\frac{1}{2}}(\hat{\beta}_{gS} - \beta) \rightarrow N(0, \Sigma_S)$ in distribution as $K \rightarrow \infty$, where Σ_S is finite and positive definite. Further, $K\widehat{V}_{gS}$ is consistent for Σ_S as $K \rightarrow \infty$, that is $K\widehat{V}_{gS} - \Sigma_S \rightarrow 0$ in probability.

4.4 Combined Estimating Equation

We now turn to the method of combined estimating equation. The method of resampling used here is the same as in Section 4.3, namely, select one unit at random from each cluster. We then combine the g equations in (4.3.11) before solving for β . This leads to

$$U_{gc}(\beta) = \frac{1}{g} \sum_{j=1}^g U_j^*(\beta) = \frac{1}{gK} \sum_{j=1}^g \sum_{i=1}^K u_{ij}^*(\beta) = 0. \quad (4.4.15)$$

In general, we solve (4.4.15) using the Newton-Raphson iterations (4.2.4) with $U_I(\hat{\beta}^{(r-1)})$ changed to $U_{gc}(\hat{\beta}^{(r-1)})$ and $J_I(\hat{\beta}^{(r-1)})$ changed to $J_{gc}(\hat{\beta}^{(r-1)})$, where

$$J_{gc}(\beta) = -\frac{\partial U_{gc}(\beta)}{\partial \beta'} = \frac{1}{gK} \sum_{j=1}^g \sum_{i=1}^K \frac{1}{a_{ij}^*} \left(\frac{\partial \mu_{ij}^*}{\partial \beta} \right)' \left(\frac{\partial \mu_{ij}^*}{\partial \beta} \right). \quad (4.4.16)$$

At convergence, we get the estimator $\hat{\beta}_{gc}$ as well as $J_{gc}(\hat{\beta}_{gc})$, the observed information matrix divided by the number of clusters, K . The variance estimator of $\hat{\beta}_{gc}$ is given by

$$\widehat{V}(\hat{\beta}_{gc}) = \{J_{gc}(\hat{\beta}_{gc})\}^{-1} \left[\frac{1}{g} \sum_{j=1}^g \left\{ \frac{1}{K^2} \sum_{i \in s_j^*} u_{ij}^*(\hat{\beta}_{gc}) u_{ij}^*(\hat{\beta}_{gc})' - U_j^*(\hat{\beta}_{gc}) U_j^*(\hat{\beta}_{gc})' \right\} \right] \times \{J_{gc}(\hat{\beta}_{gc})\}^{-1}. \quad (4.4.17)$$

The combined estimating equation method is computationally simpler than the Hoffman et al. (2001) method because we need to solve only one estimating equation (4.4.15) to get $\hat{\beta}_{gc}$ of β . Some properties of the combined estimating function $U_{gc}(\beta)$ are given in Theorem 4.1 below.

Theorem 4.1 1. Conditional on the original sample s , $U_{gc}(\beta)$ converges to

$E\{U_{gc}(\beta)|s\} = U_{\infty}(\beta)$, as $g \rightarrow \infty$, where

$$U_{\infty}(\beta) = \frac{1}{K} \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)' \frac{(y_{ij} - \mu_{ij})}{a_{ij}} = \frac{1}{K} \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} u_{ij}(\beta). \quad (4.4.18)$$

2. $E\{U_{gc}(\beta)\} = E\{U_1^*(\beta)\}$.

3. $\text{var}\{U_{gc}(\beta)\} = \text{var}\{U_{\infty}(\beta)\} + \frac{1}{g} E[\text{var}\{U_1^*(\beta)|s\}]$.

4. The variance of $U_{gc}(\beta)$ may alternatively be written as

$$\text{var}\{U_{gc}(\beta)\} = \text{var}\{U_1^*(\beta)\} - \frac{g-1}{g} E[\text{var}\{U_1^*(\beta)|s\}].$$

Proof

1. The result follows by noting that conditional on s , $U_1^*, \dots, U_g^*$ are independent identically distributed bounded random vectors.

2. This follows from the standard relationship between conditional and unconditional expectations:

$$E\{U_{gc}(\beta)\} = E[E\{U_1^*(\beta)|s\}] = E\{U_1^*(\beta)\}.$$

3. This follows from the corresponding results for variances, and the conditional independence of the U_j^* 's given s :

$$\begin{aligned} \text{var}\{U_{gc}(\beta)\} &= \text{var}[E\{U_{gc}(\beta)|s\}] + E[\text{var}\{U_{gc}(\beta)|s\}] \\ &= \text{var}\{U_{\infty}(\beta)\} + \frac{1}{g} E[\text{var}\{U_1^*(\beta)|s\}]. \end{aligned}$$

4. Apply Result 3 of Theorem 4.1 with $g = 1$ to obtain

$$\text{var} \{U_\infty(\beta)\} = \text{var} \{U_1^*(\beta)\} - E[\text{var} \{U_1^*(\beta)|s\}],$$

and substituting into Result 3 of Theorem 4.1, the result follows. ■

Theorem 4.2 *Let $\hat{\beta}_\infty$ be the solution to $U_\infty(\beta) = 0$ and $\hat{\beta}_{gc}$ the solution to $U_{gc}(\beta) = 0$. Then under mild regularity conditions, $\hat{\beta}_{gc}$ and $\hat{\beta}_\infty$ are asymptotically equivalent and furthermore have the same asymptotic distribution, as $g \rightarrow \infty$ and $K \rightarrow \infty$.*

Proof

To establish the proof of Theorem 4.2, we make the following assumptions

A1. $\sup_i E \left\| \frac{\partial u_{ij}(\beta)}{\partial \beta} \right\|^{1+\alpha} < \infty$ for some $\alpha > 0$

A2. By assumption (A1) and the law of large numbers $-\partial U_\infty(\beta)/\partial \beta'$ converges in probability to $I_\infty^*(\beta)$ as $K \rightarrow \infty$, where

$$I_\infty^*(\beta) = \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{i=1}^K E \left\{ -\frac{\partial}{\partial \beta'} \frac{1}{n_i} \sum_{j=1}^{n_i} u_{ij}(\beta) \right\}$$

is assumed to exist and it is nonsingular.

Under the assumptions A1, A2 and using Taylor expansion of $U_\infty(\hat{\beta}_\infty)$ around β we have

$$\hat{\beta}_\infty - \beta = \{I_\infty^*(\beta)\}^{-1} U_\infty(\beta) + o_p(1/\sqrt{K}). \quad (4.4.19)$$

Similarly,

$$\hat{\beta}_{gc} - \beta = \{I_\infty^*(\beta)\}^{-1} U_{gc}(\beta) + o_p(1/\sqrt{K}). \quad (4.4.20)$$

From (4.4.19) and (4.4.20)

$$\|\hat{\beta}_{\infty} - \hat{\beta}_{gc}\| \leq \|\{I_{\infty}^*(\beta)\}^{-1}\| \left\| \frac{1}{gK} \sum_{j=1}^g \sum_{i \in s_j^*} u_{ij}(\beta) - \frac{1}{K} \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} u_{ij}(\beta) \right\| + o_p(1/\sqrt{K}).$$

Apply the law of large numbers to the resampling distribution, the first term converges to zero in probability as $g \rightarrow \infty$. Hence, as $g \rightarrow \infty$ we get $\hat{\beta}_{\infty} - \hat{\beta}_{gc} = o_p(1/\sqrt{K})$. Thus $\hat{\beta}_{\infty}$ and $\hat{\beta}_{gc}$ are asymptotically equivalent as $K \rightarrow \infty$. Furthermore, it follows that $\hat{\beta}_{\infty}$ and $\hat{\beta}_{gc}$ have the same asymptotic distribution. ■

4.5 Mean Estimating Equation

From Result 1 of Theorem 4.1 we can avoid repeated resampling by using the mean estimating function $U_{\infty}(\beta) =: U_M(\beta)$ given by (4.4.18) instead of the combined estimating function $U_{gc}(\beta)$. But note that $U_M(\beta)$ requires cluster identifiers, unlike $U_{gc}(\beta)$ or $U_j^*(\beta), j = 1, \dots, g$. In biological and biomedical applications, cluster identifiers are often provided in the data file, but in social sciences applications, cluster identifiers are not revealed due to confidentiality reasons. In the latter case, the combined estimating equation method can be implemented, provided the public-use data file consists of multiple subsamples $s_1^*, \dots, s_g^*$ without cluster identifiers, and randomly permuted data vectors in each subsample. Note that $U_{gc}(\beta)$ and the estimated covariance matrix of $\hat{\beta}_{gc}$, given by (4.4.15) and (4.4.17) respectively are

invariant to permutations within each subsample.

The mean estimating equation (MEE) is given by

$$U_M(\beta) = \frac{1}{K} \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)' \frac{(y_{ij} - \mu_{ij})}{a_{ij}} = \frac{1}{K} \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\beta) = \mathbf{0}. \quad (4.5.21)$$

We solve (4.5.21) using the Newton-Raphson iterations (4.2.4) with $U_I(\hat{\beta}^{(r-1)})$ changed to $U_M(\hat{\beta}^{(r-1)})$ and $J_I(\hat{\beta}^{(r-1)})$ changed to $J_M(\hat{\beta}^{(r-1)})$, where $U_M(\hat{\beta}^{(r-1)})$ and $J_M(\hat{\beta}^{(r-1)})$ are the values of $U_M(\beta)$ and

$$J_M(\beta) = -\frac{\partial U_M(\beta)}{\partial \beta'} = \frac{1}{K} \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{a_i} \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)' \left(\frac{\partial \mu_{ij}}{\partial \beta} \right) \quad (4.5.22)$$

evaluated at $\beta = \hat{\beta}^{(r-1)}$. At convergence, we get the estimator $\hat{\beta}_M$ as well as $J_M(\hat{\beta}_M)$, the observed information matrix divided by the number of clusters, K . The variance estimator of $\hat{\beta}_M$ is given by

$$\widehat{V}_M = \{J_M(\hat{\beta}_M)\}^{-1} \left[\frac{1}{K^2} \sum_{i=1}^K \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\hat{\beta}_M) \right\} \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\hat{\beta}_M) \right\}' \right] \times \{J_M(\hat{\beta}_M)\}^{-1}. \quad (4.5.23)$$

Note that $\widehat{V}_M$ is positive definite unlike $\widehat{V}_{gs}$ or $\widehat{V}_{gc}$ so that the variance estimators of the components of $\hat{\beta}_M$ are always positive.

The MEE method can be implemented easily in any software package (such as SUDAAN or STATA) that can handle survey designs. The precise syntax varies from package to package, but in all of them we need only specify a vector giving the cluster to which each belongs and a vector of case weights with components equal to the inverse of the corresponding cluster size.

A paper on CEE and MEE was submitted to *Biometrika* in September 2002, and while revising that paper, Williamson, Datta and Satten (2003) published a paper

which also proposed the mean estimating equation method. They established the asymptotic normality of $K^{\frac{1}{2}}(\hat{\beta}_M - \beta)$ as $K \rightarrow \infty$ by proving that $K^{\frac{1}{2}}(\hat{\beta}_M - \beta) \rightarrow N(0, \Sigma_M)$, where Σ_M is finite and positive definite. Further, $K\widehat{V}_M$ is consistent for Σ_M as $K \rightarrow \infty$, that is $K\widehat{V}_M - \Sigma_M \rightarrow 0$ in probability. In this thesis, we therefore study some other properties of the mean estimating equation method not covered by Williamson et al. (2003).

4.6 Efficiency of the Mean Estimating Equation Estimator

In this section we study the efficiency of the mean estimating equation estimator, $\hat{\beta}_M$, relative to the Liang-Zeger estimator, $\hat{\beta}_I$, under both ignorable and nonignorable cluster sizes for the three examples studied in Section 4.2.

4.6.1 Mean Model

Consider the mean model (4.2.5). The Liang-Zeger independence estimating equations estimator of μ , $\hat{\mu}_I$ is given by (4.2.6) and the MEE estimator is

$$\hat{\mu}_M = K^{-1} \sum_{i=1}^K \bar{y}_i,$$

where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$. The MEE estimator, $\hat{\mu}_M$ is unbiased but the IEE estimator, $\hat{\mu}_I \rightarrow_p \tilde{\mu} = \mu + \bar{n}^{-1} \text{cov}(n_i, \bar{y}_i)$. Thus $\hat{\mu}_I$ is asymptotically biased for μ unless the n_i 's and y_{ij} 's are uncorrelated.

For ignorable cluster sizes, where the n_i 's and y_{ij} 's are independent, both $\hat{\mu}_I$ and

$\hat{\mu}_M$ are unbiased for μ . It is also easy to show that

$$\text{var}(\hat{\mu}_I) \approx \frac{\sigma^2}{K\bar{n}} \left\{ 1 + (\bar{n} - 1)\rho + \rho \frac{V_n}{\bar{n}} \right\}, \quad (4.6.24)$$

and

$$\text{var}(\hat{\mu}_M) \approx \frac{\sigma^2}{K\bar{n}} \left[1 + (\bar{n} - 1)\rho + (1 - \rho) \left\{ \bar{n}E(n_i^{-1}) - 1 \right\} \right], \quad (4.6.25)$$

where σ^2 and ρ are the variance and intra-cluster correlation of the y_{ij} 's, V_n and $\bar{n}$ are the variance and the mean of the n_i 's. It follows from (4.6.24) and (4.6.25) that the relative efficiency depends on ρ and the variability of the n_i 's as measured by V_n and $E(n_i^{-1}) - \bar{n}^{-1}$, with the efficiency of $\hat{\mu}_M$ increasing as ρ increases. Note that $\hat{\mu}_I$ is always more efficient if $\rho = 0$ and $\hat{\mu}_M$ is always more efficient if $\rho = 1$. For the purposes of illustration let $\sigma_\alpha^2 = 0.25$ and $\sigma_\epsilon^2 = 1$ so that $\rho = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\epsilon^2) = 0.20$. Further, suppose that $n_i = 1 + n_i^*$, where n_i^* is a Poisson random variable with mean 4 so that $E(n_i) = 5$ and $\text{var}(n_i) = 4$. We assume that the number of clusters $K = 100$. It now follows from (4.6.24) and (4.6.25) that $\text{var}(\hat{\beta}_{0I}) \approx 4.9 \times 10^{-3}$ and $\text{var}(\hat{\beta}_{0M}) \approx 4.954 \times 10^{-3}$, noting that $E(n_i^{-1}) = 0.2454$.

For nonignorable cluster sizes n_i , we assume that n_i^* is a Poisson random variable with mean $\lambda(\alpha_i)$, where $\log \lambda(\alpha_i) = \gamma_0 + \gamma_1 \alpha_i$ with $\gamma_0 = 0.9736$ and $\gamma_1 = 1$ and α_i is normally distributed with mean 0 and variance $\sigma_\alpha^2 = 0.25$. In this case, $\hat{\mu}_M$ is unbiased but $\hat{\mu}_I$ is asymptotically biased: $E(\hat{\mu}_I) - \mu \approx \bar{n}^{-1} \text{cov}(n_i, \bar{y}_i) = 0.1875$. Further, $\text{var}(\hat{\mu}_I) \approx 5.868 \times 10^{-3}$ whereas $\text{MSE}(\hat{\mu}_I) = 41.024 \times 10^{-3}$. The large value of $\text{MSE}(\hat{\mu}_I)$ is due to the large bias of $\hat{\mu}_I$ relative to the standard error. The expression (4.6.25) for $\text{var}(\hat{\mu}_M)$ involves $E(n_i^{-1})$ which cannot be evaluated analytically. We

present simulation results in Section 4.8 on the efficiency of $\hat{\mu}_M$ for the mean model (4.2.5) under nonignorable cluster sizes.

4.6.2 Linear regression model

Consider the nested error linear regression model (4.2.7) with $\mathbf{x}_i = (1, z_i)'$, where z_i is a cluster-specific covariate taking the values 0 or 1. Let $\mu_l = E(y_{ij}|z_i = l)$, $l = 0, 1$. We can then write $\beta_0 = \mu_0$ and $\beta_1 = \mu_1 - \mu_0$. The independence estimating equations estimators of μ_0 and μ_1 are

$$\hat{\mu}_{lI} = \left(\sum_{S_l} n_i \right)^{-1} \left(\sum_{S_l} n_i \bar{y}_i \right), l = 0, 1$$

where S_l denotes the set of sample clusters with $z_i = l$ for $l = 0, 1$. Hence the independence estimating equations estimators of β_0 and β_1 are given by

$$\hat{\beta}_{0I} = \hat{\mu}_{0I}, \hat{\beta}_{1I} = \hat{\mu}_{1I} - \hat{\mu}_{0I}.$$

The mean estimating equation estimators of μ_0 and μ_1 , are

$$\hat{\mu}_{lM} = \frac{1}{K_l} \sum_{S_l} \bar{y}_i, l = 0, 1$$

where K_l is the size of S_l . Hence, the mean estimating equation estimators of β_0 and β_1 are given by

$$\hat{\beta}_{0M} = \hat{\mu}_{0M}, \hat{\beta}_{1M} = \hat{\mu}_{1M} - \hat{\mu}_{0M}.$$

The estimators $\hat{\beta}_{0M}$ and $\hat{\beta}_{1M}$ are unbiased. On the other hand, $E(\hat{\mu}_{lI}) \approx \mu_l + \bar{n}_l^{-1} \text{cov}_l(n_i, \bar{y}_i)$ so that $\hat{\mu}_{lI}$ is biased for μ_l if n_i 's and y_{ij} 's are correlated given $z_i = l$.

For the specification $n_i = h(\alpha_i)$ in Example 4.2, $\bar{n}_l$ and $\text{cov}_l(n_i, \bar{y}_i)$ do not depend on l so that $\hat{\beta}_{1I}$ is unbiased while $\hat{\beta}_{0I}$ is asymptotically biased.

For ignorable cluster sizes, we assume that the n_i 's and y_{ij} 's are distributed independently. Then both $\hat{\mu}_{lI}$ and $\hat{\mu}_{lM}$ are unbiased for μ_l . It is easy to show that

$$\text{var}(\hat{\mu}_{lI}) \approx \frac{\sigma_l^2}{\bar{n}K_l} \left\{ 1 + (\bar{n} - 1)\rho_l + \rho_l \frac{V_n}{\bar{n}} \right\} \quad (4.6.26)$$

and

$$\text{var}(\hat{\mu}_{lM}) \approx \frac{\sigma_l^2}{\bar{n}K_l} \left[1 + (\bar{n} - 1)\rho_l + (1 - \rho_l) \left\{ \bar{n}E(n_i^{-1}) - 1 \right\} \right], \quad (4.6.27)$$

where σ_l^2 and ρ_l are the conditional variance and the intra-cluster correlation of the y_{ij} 's given $z_i = l$, and $\bar{n}$ and V_n are the mean and variance of the n_i 's. It follows from (4.6.26) and (4.6.27) that the relative efficiency depends on ρ_l and the variability of n_i 's as measured by V_n and $E(n_i^{-1}) - \bar{n}^{-1}$, with the efficiency of $\hat{\mu}_{lM}$ increasing as ρ_l increases. As in the case of the mean model, $\hat{\mu}_{lI}$ is always more efficient if $\rho_l = 0$, and $\hat{\mu}_{lM}$ is always more efficient if $\rho_l = 1$.

For nonignorable cluster sizes, we present simulation results in Section 4.8 on the efficiency of $\hat{\beta}_M$ relative to $\hat{\beta}_I$.

4.6.3 Logistic linear regression

We now consider the logistic linear regression model (4.2.10) with cluster-specific covariates z_i taking values 0 or 1. In this case, $\text{var}(\hat{\pi}_{lI})$ and $\text{var}(\hat{\pi}_{lM})$ are given by (4.6.26) and (4.6.27). Hence, the results on $\text{var}(\hat{\mu}_{lI})$ and $\text{var}(\hat{\mu}_{lM})$ for linear regression also hold for $\text{var}(\hat{\pi}_{lI})$ and $\text{var}(\hat{\pi}_{lM})$. We present simulation results in Section 4.8 on

the efficiency of $\hat{\beta}_M$ for the logistic linear regression model (4.2.10) with nonignorable cluster sizes.

4.7 Wald tests for Ignorability of Cluster Sizes

As noted in Section 4.6, the mean estimating equation estimator $\hat{\beta}_M$ is valid for both nonignorable and ignorable cluster sizes. However, when cluster sizes are ignorable, the independence estimating equations estimator $\hat{\beta}_I$ may be more efficient. Further, as noted in Subsection 4.6.2, the slope component of $\hat{\beta}_I$ may not be biased when cluster sizes are nonignorable. It is therefore useful to check for nonignorability from the data. An informal approach may be done by estimating the covariance between the cluster size n_i and the cluster mean $\bar{y}_i$ for the exposed and nonexposed groups as demonstrated in Section 4.6. However, it is also useful to have formal tests for ignorability of cluster sizes.

We now provide such a test for ignorability of cluster sizes by comparing the components of estimators $\hat{\beta}_M$ and $\hat{\beta}_I$ under the null hypothesis that cluster sizes are ignorable using a Wald test based on the difference $\hat{\beta}_M - \hat{\beta}_I$. If the null hypothesis is not tenable, it may be of interest to test whether the slope estimator $\hat{\beta}_{1I}$ is affected by nonignorability, using a Wald test based on $\hat{\beta}_{1M} - \hat{\beta}_{1I}$. If the test reveals that $\hat{\beta}_{1I}$ is not affected and if the interest is only on the slope β_1 , then one could simply use $\hat{\beta}_{1I}$ since $\hat{\beta}_{1I}$ may be more efficient than $\hat{\beta}_{1M}$ when $\hat{\beta}_{1I}$ is not biased.

Under the assumption (B1) below and using the Taylor's expansion of $U_I(\hat{\beta}_I)$

around β to get

$$\hat{\beta}_I - \beta = \{I_I^*(\beta)\}^{-1} \left\{ \frac{1}{N} \sum_{i=1}^K \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\beta) \right\} + o_p(1/\sqrt{K}), \quad (4.7.28)$$

where $U_I(\beta)$ is the Liang-Zeger independence estimating function in (4.2.3).

Now apply assumption (B1) to $-\partial U_M(\beta)/\partial \beta'$ and the Taylor's expansion of $U_M(\hat{\beta}_M)$ around β leads to

$$\hat{\beta}_M - \beta = \{I_M^*(\beta)\}^{-1} \left\{ \frac{1}{K} \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\beta) \right\} + o_p(1/\sqrt{K}), \quad (4.7.29)$$

where $I_M^*(\beta) = \lim_{K \rightarrow \infty} E(-\partial U_M(\beta)/K \partial \beta')$ and $U_M(\beta)$ is the mean estimating function in (4.5.21). It follows from (4.7.28) and (4.7.29) that

$$\hat{\beta}_M - \hat{\beta}_I = \frac{1}{K} \sum_{i=1}^K \mathbf{u}_i^*(\beta) + o_p(1/\sqrt{K}), \quad (4.7.30)$$

where

$$\mathbf{u}_i^*(\beta) = \left[\frac{1}{n_i} \{I_M^*(\beta)\}^{-1} - \frac{K}{N} \{I_I^*(\beta)\}^{-1} \right] \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\beta). \quad (4.7.31)$$

Theorem 4.3 establishes a central limit theorem for $\hat{\beta}_M - \hat{\beta}_I$ under the null hypothesis.

Theorem 4.3 *For the null hypothesis that cluster sizes are ignorable, and under mild regularity conditions, $K^{\frac{1}{2}}(\hat{\beta}_M - \hat{\beta}_I)$ is normally distributed with mean 0 and covariance matrix estimated by*

$$\widehat{V}^* = \frac{1}{K(K-1)} \sum_{i=1}^K \left\{ \mathbf{u}_i^*(\hat{\beta}) - \bar{\mathbf{u}}^*(\hat{\beta}) \right\} \left\{ \mathbf{u}_i^*(\hat{\beta}) - \bar{\mathbf{u}}^*(\hat{\beta}) \right\}' \quad (4.7.32)$$

where $\bar{\mathbf{u}}^*(\hat{\beta}) = K^{-1} \sum_{i=1}^K \mathbf{u}_i^*(\hat{\beta})$, $\mathbf{u}_i^*(\hat{\beta})$ is given in (4.7.31), and $\hat{\beta}$ is an estimator of β which may be taken as $\hat{\beta} = \hat{\beta}_I$ or $\hat{\beta} = \hat{\beta}_M$ or $\hat{\beta} = \frac{1}{2}(\hat{\beta}_I + \hat{\beta}_M)$.

Proof

To establish the proof, we make the following assumptions:

- B1. By assumption (A1) given in the proof of Theorem 4.2 and the law of large numbers, $-\partial U_I(\beta)/\partial \beta'$ converges in probability to $I_I^*(\beta)$ as $K \rightarrow \infty$, where

$$I_I^*(\beta) = \lim_{K \rightarrow \infty} \frac{1}{N} \sum_{i=1}^K E \left\{ -\frac{\partial}{\partial \beta'} \sum_{j=1}^{n_i} u_{ij}(\beta) \right\}$$

is assumed to exist and it is nonsingular. Note that $N = \sum_{i=1}^K n_i$.

- B2. $\sup_i E \|u_{ij}(\beta)\|^{2+\alpha} < \infty$ for some $\alpha > 0$.

- B3. $K^{-1} \sum_{i=1}^K \text{var}\{u_i^*(\beta)\} \rightarrow_p \Sigma^*$ as $K \rightarrow \infty$.

Using assumptions (B2) and (B3), then for any vector a ,

$$\frac{\sum_{i=1}^K E |a' u_i^*(\beta)|^{2+\alpha}}{\left[\sum_{i=1}^K \text{Var} \{a' u_i^*(\beta)\} \right]^{\frac{2+\alpha}{2}}} = O(K^{-\frac{\alpha}{2}}) \rightarrow 0.$$

Hence by Lyapunov's central limit theorem,

$$\frac{1}{\sqrt{K}} \sum_{i=1}^K u_i^*(\beta) \rightarrow_d N(0, \Sigma^*) \text{ as } K \rightarrow \infty.$$

The result follows immediately from (4.7.30). ■

It now follows from Theorem 4.3 that a Wald statistic for testing the null hypothesis for ignorability of cluster sizes is given by

$$W = (\hat{\beta}_M - \hat{\beta}_I)' \widehat{V}^{*-1} (\hat{\beta}_M - \hat{\beta}_I), \quad (4.7.33)$$

where $\widehat{V}^*$ is given by (4.7.32). Under the null hypothesis W is asymptotically χ_p^2 , a χ^2 variable with p degrees of freedom.

Under the null hypothesis, for the special case of the mean model, $\mu_{ij} = \mu$, $a_{ij} = \sigma^2$, $u_{ij}(\mu) = (y_{ij} - \mu)$, $J_I(\mu) = 1 = I_I^*(\mu)$ and $J_M(\mu) = 1 = I_M^*(\mu)$. The representation (4.7.30) is exact. For the linear regression model with cluster-specific covariates $\mathbf{x}_i$, we have $\mu_{ij} = \mathbf{x}_i' \boldsymbol{\beta}$, $a_{ij} = \sigma^2$, $\mathbf{u}_{ij}(\boldsymbol{\beta}) = \mathbf{x}_{ij}(y_{ij} - \mathbf{x}_{ij}' \boldsymbol{\beta})$ and

$$J_I(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^K n_i \mathbf{x}_i \mathbf{x}_i' = \mathbf{I}_I^*(\boldsymbol{\beta}), \quad J_M(\boldsymbol{\beta}) = \frac{1}{K} \sum_{i=1}^K \mathbf{x}_i \mathbf{x}_i' = \mathbf{I}_M^*(\boldsymbol{\beta}). \quad (4.7.34)$$

Also, the representation (4.7.30) is exact. Similarly, for the special case of logistic linear regression with cluster-specific covariate $\mathbf{x}_i$, we have $\text{logit}(\mu_{ij}) = \mathbf{x}_i' \boldsymbol{\beta}$, $a_{ij} = \mu_{ij}(1 - \mu_{ij})$, $\mathbf{u}_{ij}(\boldsymbol{\beta}) = \mathbf{x}_i(y_{ij} - \mu_{ij})$, $\mu_{ij} = \mu_i$,

$$J_I(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^K n_i \mathbf{x}_i \mathbf{x}_i' \mu_i (1 - \mu_i) = \mathbf{I}_I^*(\boldsymbol{\beta}) \quad (4.7.35)$$

and

$$J_M(\boldsymbol{\beta}) = \frac{1}{K} \sum_{i=1}^K \mathbf{x}_i \mathbf{x}_i' \mu_i (1 - \mu_i) = \mathbf{I}_M^*(\boldsymbol{\beta}). \quad (4.7.36)$$

In Section 4.8, we present simulation results on the size and power of the Wald test based on (4.7.33) for the above three special cases.

4.8 Simulation Study

4.8.1 Mean Model

For the mean model considered in (4.2.5), we used $\sigma_\alpha^2 = 0.25$ and $\sigma_\epsilon^2 = 1$ so that $\rho = 0.20$. For the ignorable cluster sizes, we generated the cluster sizes n_i by letting

$n_i = 1 + n_i^*$, where the n_i^* 's are independent Poisson variables with mean 4 so that $E(n_i) = 5$ and $\text{var}(n_i) = 4$. For $K = 100$ and a realized n_i -pattern, the responses y_{ij} were generated from the model (4.2.5) with $\mu = -0.20$.

For the nonignorable cluster sizes, the n_i^* 's were generated from a Poisson distribution with mean $\lambda(\alpha_i)$, where $\log\{\lambda(\alpha_i)\} = \gamma_0 + \gamma_1\alpha_i$, with $\gamma_0 = 0.9736$ and $\gamma_1 = 1$. We have $E(n_i) = \bar{n} = 1 + E(n_i^*) = 1 + E(e^{\gamma_0 + \gamma_1\alpha_i}) = 1 + e^{\gamma_0 + \frac{1}{2}\gamma_1^2\sigma_\alpha^2} = 4$. We used $R = 1000$ independent simulation runs for each case.

(a) *Wald tests for ignorability of cluster sizes*

Table 4.1 presents the size and power of the Wald tests for ignorability of cluster sizes, (4.7.33), for nominal level $\alpha = 0.05$, using the covariance matrix (4.7.32) with (i) $\hat{\mu} = \hat{\mu}_I$, (ii) $\hat{\mu} = \hat{\mu}_M$, (iii) $\hat{\mu} = \frac{1}{2}(\hat{\mu}_I + \hat{\mu}_M)$. It is clear from Table 4.1 that the proposed test (4.7.33) performs very well in terms of both size and power regardless of the choice of $\hat{\mu}$ in (4.7.32).

Table 4.1: *Size and power of the Wald tests for ignorability of cluster sizes based on 1000 simulations: mean model, nominal level test=0.05 and $K = 100$ clusters.*

Estimate of μ used in $\hat{V}^*$	Size of Wald test (standard error)	Power of Wald test (standard error)
IEE $\hat{\mu}_I$	0.054 (0.0071)	0.990 (0.0031)
MEE $\hat{\mu}_M$	0.054 (0.0071)	0.988 (0.0034)
$\frac{1}{2}(\hat{\mu}_I + \hat{\mu}_M)$	0.051 (0.0071)	0.989 (0.0033)

IEE, independence estimating equations; MEE, mean estimating equation.

(b) *Performance of estimators of μ*

Table 4.2 reports the means and mean squared errors of the estimators $\hat{\mu}_M$, $\hat{\mu}_I$, $\hat{\mu}_{gS}$ and $\hat{\mu}_{gc}$, relative biases of the variance estimators and the coverage probabilities of normal theory intervals based on $(\hat{\mu}_M, \hat{V}_M)$, $(\hat{\mu}_I, \hat{V}_I)$, $(\hat{\mu}_{gS}, \hat{V}_{gS})$ and $(\hat{\mu}_{gc}, \hat{V}_{gc})$. The relative bias of a variance estimator is given by $\{E(\text{var.est.}) - \text{MSE}(\text{est.})\} / \text{MSE}(\text{est.})$. It shows that all the estimators $\hat{\mu}_M$, $\hat{\mu}_I$, $\hat{\mu}_{gS}$ and $\hat{\mu}_{gc}$ are unbiased for the ignorable case, while $\hat{\mu}_I$ is heavily biased in the case of nonignorable cluster sizes: mean = -0.0128 compared to the true $\beta_0 = -0.20$. On the other hand, $\hat{\mu}_M$, $\hat{\mu}_{gS}$ and $\hat{\mu}_{gc}$ are unbiased in the latter case. For the ignorable case, comparing the mean squared errors in Table 4.2, it follows that the estimators $\hat{\mu}_I$, $\hat{\mu}_M$, $\hat{\mu}_{gS}$ and $\hat{\mu}_{gc}$ are about equally efficient. However, for the nonignorable cluster size case, $\hat{\mu}_I$ has a much larger mean squared error than any of the estimators due to large bias: simulated bias of 0.1872 compared to the asymptotic bias, $\bar{n}^{-1} \text{Cov}(n_i, \bar{y}_i) = \frac{1}{4} \left\{ \sigma_\alpha^2 \gamma_1 e^{\gamma_0 + \frac{1}{2} \gamma_1^2 \sigma_\alpha^2} \right\} = \frac{1}{4}(0.75) = 0.1875$.

For the ignorable cluster size case, the relative bias of the variance estimator is less than 10% for all three methods. On the other hand, for the nonignorable cluster size case, the variance estimator of $\hat{\mu}_I$ leads to serious underestimation: -84.60%.

Turning to coverage probabilities of 0.95-level normal theory intervals, all the methods perform well in the ignorable case. On the other hand, for the case of nonignorable cluster sizes, $\hat{\mu}_I$ leads to serious undercoverage: 0.343 versus nominal 0.950.

4.8.2 Linear regression

For the special case of linear regression, we considered the model (4.2.7) with a

Table 4.2: Means and mean squared errors of estimators of μ , relative biases of variance estimators and coverage probabilities of associated 0.95-level normal theory intervals based on 1000 simulations: mean model with $\mu = -0.2$ and $K = 100$ clusters.

	MEE	CEE	SEE	IEE
<i>Ignorable cluster sizes</i>				
E(est.)	-0.2011	-0.2021	-0.2021	-0.1984
μ : MSE(est.) $\times 1000$	4.9375	5.2041	5.2041	4.8154
RB(var.est.) $\times 100$	-6.6500	-5.1310	-6.5800	-4.5000
Coverage probability	0.937	0.941	0.940	0.943
<i>Nonignorable cluster sizes</i>				
E(est.)	-0.1998	-0.1999	-0.1999	-0.0128
μ : MSE(est.) $\times 1000$	6.1700	6.1708	6.1708	41.6484
RB(var.est.) $\times 100$	-1.4200	-1.1164	-2.1982	-84.6000
Coverage probability	0.949	0.948	0.947	0.343

IEE, independence estimating equations; MEE, mean estimating equation; CEE, combined estimating equation; SEE, separate estimating equations; MSE, mean squared error; RB, relative bias

single cluster-specific binary covariate z_i , i.e., $\mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 z_i$ in (4.2.7). We used $\sigma_\alpha^2 = 0.25$ and $\sigma_\epsilon^2 = 1$ so that $\rho = 0.20$ as in Subsection 4.8.1. Further, the covariate z_i was generated by setting $z_i = 1$ if $U_i \leq 0.5$ and $z_i = 0$ otherwise, where the U_i 's are independent and identically distributed uniform random variables on the unit

interval $[0, 1]$. For the case of ignorable cluster sizes, we generated the cluster sizes n_i as in Subsection 4.8.1. For $K = 100$ and a realized n_i -pattern, the responses y_{ij} were generated from the above specification for $\beta_0 = 0.4$ and $\beta_1 = -0.2$.

For the case of nonignorable cluster sizes, the n_i^* 's were generated as in Subsection 4.8.1. We used $R = 1000$ independent simulation runs for each case.

(a) *Wald tests for ignorability of cluster sizes*

Table 4.3 presents the size and power of the Wald tests for ignorability of cluster sizes, (4.7.33), for nominal level $\alpha = 0.05$, using the covariance matrix (4.7.32) with (i) $\hat{\beta} = \hat{\beta}_I$, (ii) $\hat{\beta} = \hat{\beta}_M$, (iii) $\hat{\beta} = \frac{1}{2}(\hat{\beta}_I + \hat{\beta}_M)$. It is clear from Table 4.3 that the proposed test (4.7.33) performs very well in terms of both size and power regardless of the choice of $\hat{\beta}$ in (4.7.32). We also note that the power of the slope test is close to the size because for the model (4.2.7), the IEE slope estimator, $\hat{\beta}_{1I}$ is unaffected by nonignorability.

(b) *Performance of estimators of β*

Table 4.4 reports the means and mean squared errors of the estimators $\hat{\beta}_M$, $\hat{\beta}_{gS}$, $\hat{\beta}_{gc}$ and $\hat{\beta}_I$ and the coverage probabilities of normal theory intervals based on $(\hat{\beta}_M, \widehat{V}_M)$, $(\hat{\beta}_{gS}, \widehat{V}_{gS})$, $(\hat{\beta}_{gc}, \widehat{V}_{gc})$ and $(\hat{\beta}_I, \widehat{V}_I)$. It shows that all the estimators $\hat{\beta}_M$, $\hat{\beta}_{gS}$, $\hat{\beta}_{gc}$ and $\hat{\beta}_I$ are unbiased for the case of ignorable cluster sizes, while $\hat{\beta}_{0I}$ is heavily biased in the case of nonignorable cluster sizes: mean=0.586 compared to the true $\beta_0 = 0.40$. On the other hand, $(\hat{\beta}_{0M}, \hat{\beta}_{1M})$, $(\hat{\beta}_{0gS}, \hat{\beta}_{1gS})$ and $(\hat{\beta}_{0gc}, \hat{\beta}_{1gc})$ are unbiased in the latter case; $\hat{\beta}_{1I}$ is also unbiased. For the ignorable case, comparing the mean squared errors in Table 4.4, it follows that the estimators $(\hat{\beta}_{0I}, \hat{\beta}_{1I})$ are slightly more efficient than $(\hat{\beta}_{0M}, \hat{\beta}_{1M})$ or $(\hat{\beta}_{0gS}, \hat{\beta}_{1gS})$ or $(\hat{\beta}_{0gc}, \hat{\beta}_{1gc})$. However, for the nonignorable cluster

Table 4.3: *Size and power of the Wald tests for ignorability of cluster sizes based on 1000 simulations: linear regression, nominal level test=0.05 and $K = 100$ clusters.*

	Estimate of β used in $\widehat{V}^*$	Size of Wald test (standard error)	Power of Wald test (standard error)
IEE:	$\hat{\beta}_I = (\hat{\beta}_{0I}, \hat{\beta}_{1I})'$	0.045 (0.0066)	0.951 (0.0068)
	$\hat{\beta}_{0I}$	0.052 (0.0070)	0.822 (0.0121)
	$\hat{\beta}_{1I}$	0.067 (0.0079)	0.050 (0.0069)
MEE:	$\hat{\beta}_M = (\hat{\beta}_{0M}, \hat{\beta}_{1M})'$	0.047 (0.0067)	0.937 (0.0079)
	$\hat{\beta}_{0M}$	0.056 (0.0073)	0.790 (0.0128)
	$\hat{\beta}_{1M}$	0.068 (0.0080)	0.040 (0.0062)
IEE and MEE:	$\frac{1}{2}(\hat{\beta}_I + \hat{\beta}_M)$	0.047 (0.0067)	0.944 (0.0073)
	$\frac{1}{2}(\hat{\beta}_{0I} + \hat{\beta}_{0M})$	0.055 (0.0072)	0.812 (0.0124)
	$\frac{1}{2}(\hat{\beta}_{1I} + \hat{\beta}_{1M})$	0.068 (0.0080)	0.044 (0.0065)

IEE, independence estimating equations; MEE, mean estimating equation.

size case, $\hat{\beta}_{0I}$ has a much larger mean squared error than either $\hat{\beta}_{0M}$ or $\hat{\beta}_{0gS}$ or $\hat{\beta}_{0gc}$ due to large bias: simulated bias of 0.1860 compared to the asymptotic bias, $\bar{n}^{-1}\text{Cov}(n_i, \bar{y}_i) = \frac{1}{4} \left\{ \sigma_\alpha^2 \gamma_1 e^{\gamma_0 + \frac{1}{2} \gamma_1^2 \sigma_\alpha^2} \right\} = \frac{1}{4}(0.75) = 0.1875$. Further, $\hat{\beta}_{1I}$ has a slightly larger mean squared error than $\hat{\beta}_{1M}$ or $\hat{\beta}_{1gS}$ or $\hat{\beta}_{1gc}$ although all are unbiased here.

Turning to coverage probabilities of 0.95-level normal theory intervals, all the methods perform well in the ignorable case. On the other hand, for the case of nonignorable cluster sizes, $\hat{\beta}_{0I}$ leads to serious undercoverage: 0.643 versus nominal 0.950. All the estimators $\hat{\beta}_{1I}$, $\hat{\beta}_{1M}$, $\hat{\beta}_{1gS}$ and $\hat{\beta}_{1gc}$ perform well in terms of coverage

Table 4.4: Means and mean squared errors of estimators of β_0 and β_1 and coverage probabilities of associated 0.95-level normal theory intervals based on 1000 simulations: linear regression with $\beta_0 = 0.4$, $\beta_1 = -0.2$, $K = 100$ clusters and $g = 1000$ repeated subsampling.

	MEE	CEE	SEE	IEE
<i>Ignorable cluster sizes</i>				
E(est.)	0.402	0.402	0.402	0.403
β_0 : MSE(est.) $\times 1000$	10.328	10.361	10.361	9.788
Coverage probability	0.935	0.931	0.929	0.945
E(est.)	-0.203	-0.203	-0.203	-0.204
β_1 : MSE(est.) $\times 1000$	20.772	20.844	20.844	18.855
Coverage probability	0.936	0.937	0.934	0.956
<i>Nonignorable cluster sizes</i>				
E(est.)	0.402	0.402	0.402	0.586
β_0 : MSE(est.) $\times 1000$	11.986	12.003	12.003	46.835
Coverage probability	0.939	0.944	0.942	0.643
E(est.)	-0.203	-0.203	-0.203	-0.202
β_1 : MSE(est.) $\times 1000$	24.907	24.918	24.918	25.728
Coverage probability	0.943	0.943	0.939	0.956

IEE, independence estimating equations; MEE, mean estimating equation; MSE, mean squared error.

probability.

4.8.3 Logistic linear regression

For the special case of logistic linear regression model (4.2.10) with binary responses y_{ij} and cluster-specific covariate z_i taking values 0 or 1, we followed Hoffman et al. (2001) to simulate samples for the ignorable and nonignorable cluster sizes cases.

Let

$$\text{logit}\{E(y_{ij}|b_i, z_i)\} = \text{logit}(b_i) + \eta z_i, \quad (4.8.37)$$

where the random effects b_i are independent and identically distributed and $\text{logit}(a) = \log\{a(1-a)^{-1}\}$. From (4.8.37) we obtain the marginal expectations $E(y_{ij}|z_i = 0) = E_0(y_{ij})$ and $E(y_{ij}|z_i = 1) = E_1(y_{ij})$ as $E_0(y_{ij}) = E(b_i)$; $E_1(y_{ij}) = E\{1 + e^{-(a_i + \eta)}\}^{-1}$, where $a_i = \text{logit}(b_i)$. Now, let $\text{logit}\{E_0(y_{ij})\} = \beta_0$ and $\text{logit}\{E_1(y_{ij})\} = \beta_0 + \beta_1$ to get the marginal model (4.2.10).

For the simulation, the covariate z_i takes the value 1 if $U_i \leq 0.5$, and the value 0 otherwise, where U_i is a random variable over the unit interval $[0, 1]$. The b_i 's are drawn from Beta ($a = 1, b = 3$) distribution leading to intraclass correlation $\rho = (a+b+1)^{-1} = 0.2$. Also, we let $\eta = 0.6111$ in (4.2.10) which leads to $\beta_0 = -1.0986$ and $\beta_1 = 0.4796$. The binary response y_{ij} takes the value 1 if $U_{ij} \leq E(y_{ij}|b_i, z_i)$ and the value 0 otherwise, where U_{ij} is a uniform variable over $[0, 1]$.

To generate ignorable cluster sizes, n_i 's were drawn from Binomial ($n = 8, p = 0.8$) distribution, truncated by discarding clusters assigned size zero. For the nonignorable cluster sizes case, we let $n_i = n(b_i)$, where $n(b_i)$ is drawn from Binomial ($n = 9, p = 0.25$) or Binomial ($n = 9, p = 0.75$) according as $b_i > a/(a+b)$ or $b_i \leq 0.25$, discarding

clusters of size 0 or 9.

We studied the performances of mean estimating equation, independence estimating equations and separate estimating equation estimators, $\hat{\beta}_M$, $\hat{\beta}_I$ and $\hat{\beta}_{gS}$. We simulated $R = 1000$ independent samples with number of clusters $K = 100$. Results for $\hat{\beta}_{gS}$ are obtained from Tables 4.17 and 5.4 of E. B. Hoffman's unpublished 1988 thesis from the University of North Carolina. She generated $R = 1000$ independent samples and $g = 10,000$ subsamples from each simulated sample.

(a) *Wald tests for ignorability of cluster sizes*

Table 4.5 presents the size and power of the Wald tests for ignorability of cluster sizes, (4.7.33) for nominal level $\alpha = 0.05$, using the estimated covariance matrix (4.7.32) with (i) $\hat{\beta} = \hat{\beta}_I$, (ii) $\hat{\beta} = \hat{\beta}_M$ and (iii) $\hat{\beta} = \frac{1}{2}(\hat{\beta}_I + \hat{\beta}_M)$. It is clear from Table 4.5 that the proposed test (4.7.33) performs very well in terms of both size and power, regardless of the selection of $\hat{\beta}$ in (4.7.32). We also note that the slope test is close to the size under (4.8.37) used in the simulation, since the slope is unaffected by nonignorability as explained in Section 4.2.

(b) *Performance of estimators of β*

Table 4.6 reports the means and mean squared errors of the estimators $\hat{\beta}_M$, $\hat{\beta}_{gS}$ and $\hat{\beta}_I$, relative biases of the variance estimators based on the diagonal elements of $\widehat{V}_M$, $\widehat{V}_{gS}$ and $\widehat{V}_I$, and the coverage probabilities of 0.95-level normal theory intervals based on $(\hat{\beta}_M, \widehat{V}_M)$, $(\hat{\beta}_{gS}, \widehat{V}_{gS})$ and $(\hat{\beta}_I, \widehat{V}_I)$.

For the ignorable case, Table 4.6 shows that all the methods perform well with respect to bias of the estimator, while $\hat{\beta}_{0I}$ is heavily biased in the case of nonignorable cluster sizes: $\hat{\beta}_{0I} = -1.510$ compared to the true $\beta_0 = -1.0986$. For the slope β_1 , all

Table 4.5: *Size and power of the Wald tests for ignorability of cluster sizes based on 1000 simulations: logistic linear regression, nominal level test=0.05 and $K = 100$ clusters.*

	Estimate of β used in $\hat{V}^*$	Size of Wald test (standard error)	Power of Wald test (standard error)
IEE	$\hat{\beta}_I = (\hat{\beta}_{0I}, \hat{\beta}_{1I})'$	0.055 (0.0072)	0.980 (0.0044)
	$\hat{\beta}_{0I}$	0.054 (0.0071)	0.882 (0.0102)
	$\hat{\beta}_{1I}$	0.056 (0.0073)	0.043 (0.0064)
MEE	$\hat{\beta}_M = (\hat{\beta}_{0M}, \hat{\beta}_{1M})'$	0.056 (0.0073)	0.982 (0.0042)
	$\hat{\beta}_{0M}$	0.057 (0.0073)	0.881 (0.0102)
	$\hat{\beta}_{1M}$	0.064 (0.0077)	0.036 (0.0059)
IEE and MEE	$\frac{1}{2}(\hat{\beta}_I + \hat{\beta}_M)$	0.056 (0.0073)	0.982 (0.0042)
	$\frac{1}{2}(\hat{\beta}_{0I} + \hat{\beta}_{0M})$	0.057 (0.0073)	0.884 (0.0101)
	$\frac{1}{2}(\hat{\beta}_{1I} + \hat{\beta}_{1M})$	0.060 (0.0075)	0.043 (0.0064)

IEE, independence estimating equations; MEE, mean estimating equation.

the methods perform well with respect to bias, thus indicating $\hat{\beta}_{1I}$ is not affected by nonignorability, as shown in Section 4.2.

Turning to mean squared errors of the estimators for the ignorable case, the estimators $(\hat{\beta}_{0M}, \hat{\beta}_{1M})$ and $(\hat{\beta}_{0I}, \hat{\beta}_{1I})$ are similar with respect to the mean squared error, while $(\hat{\beta}_{0gS}, \hat{\beta}_{1gS})$ are slightly less efficient. However, for the nonignorable cluster size case, $\hat{\beta}_{0I}$ has a much larger mean squared error than $\hat{\beta}_{0M}$ due to large bias. For the slope coefficient, $\hat{\beta}_{1I}$ is more efficient while $\hat{\beta}_{1M}$ and $\hat{\beta}_{1gS}$ have similar mean squared

errors: 0.079356 for $\hat{\beta}_{1I}$ compared to 0.106563 for $\hat{\beta}_{1M}$ and 0.106490 for $\hat{\beta}_{1gS}$. This result suggests that if the interest is only on the slope and if the Wald test indicates that $\hat{\beta}_{1I}$ is not affected by nonignorability, then it may be better to use $\hat{\beta}_{1I}$. If both β_0 and β_1 are of interest, then one could use $\hat{\beta}_{0M}$ for β_0 and $\hat{\beta}_{1I}$ for β_1 , if the Wald tests indicate that $\hat{\beta}_{0I}$ is affected by nonignorability but not $\hat{\beta}_{1I}$.

Table 4.6 shows that the relative bias of the variance estimators associated with $\hat{\beta}_{0M}$ and $\hat{\beta}_{0I}$ is less than 10% for the ignorable case. On the other hand, for the nonignorable case, the variance estimator of $\hat{\beta}_{0I}$ leads to serious underestimation: -79.488%, while the relative bias of variance estimator of $\hat{\beta}_{1M}$ is less than 10%. For the slope β_1 , the variance estimator of $\hat{\beta}_{1M}$ and $\hat{\beta}_{1I}$ both perform well with respect to relative bias.

Turning to coverage probabilities of 0.95-level normal theory intervals, all the methods perform well in the ignorable cluster size case. On the other hand, for the nonignorable cluster size case, $\hat{\beta}_{0I}$ leads to serious undercoverage: 0.510 versus nominal 0.95, whereas both $\hat{\beta}_{0M}$ and $\hat{\beta}_{0gS}$ perform well. In case of the slope parameter β_1 , all four methods perform well in terms of coverage probability.

4.8.4 Case of small K

As noted in Section 4.3, the separate estimating equations method can lead to an unstable estimated covariance matrix $\widehat{V}_{gS}$ quite often if the number of sample clusters, K , is small, even for a fairly large number of subsamples, g . In particular, the variance estimates can take negative values quite often. On the other hand, the mean

estimating equation method always gives positive variance estimates.

We conducted a simulation study for the ignorable cluster sizes case with small $K = 12$ to illustrate the instability of $\widehat{V}_{gS}$. We used a nested linear regression model given by

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \alpha_i + \epsilon_{ij}; j = 1, \dots, n_i; i = 1, \dots, K = 12, \quad (4.8.38)$$

for this purpose, where α_i and ϵ_{ij} are independent $N(0, \sigma_\alpha^2)$ and $N(0, \sigma_\epsilon^2)$, respectively, and n_i ranged from 1 to 5. The auxiliary variables (x_{1ij}, x_{2ij}) were taken from Battese, Harter & Fuller (1988), where x_{1ij} and x_{2ij} represent the number of pixels classified as corn and soybeans respectively by land observatory satellite in the j -th sample segment of the i -th county in North Central Iowa and n_i is the number of sample segments in the i -th country. The data for the second segment in Hardin county were deleted as the reported hectares of corn, y , was found to be erroneous. For this simulation study, we treated the parameter estimates corresponding to the number of hectares of corn, y , reported in Battese et al. (1988), as true values: $\beta_0 = 51$, $\beta_1 = 0.329$, $\beta_2 = -0.134$, $\sigma_\alpha^2 = 140$ and $\sigma_\epsilon^2 = 150$. Using the above specifications, we simulated $R = 1000$ independent samples $\{y_{ij}\}$ from the model (4.8.38). We generated $g = 500, 1000$ and 4000 subsamples from each sample $\{y_{ij}, x_{ij}\}$: note that the x_{ij} and n_i are fixed. Using the simulated data, variance estimates for the separate and combined estimating equations estimators of β_0 , β_1 and β_2 were computed using the diagonal elements of $\widehat{V}_{gS}$ given by (4.3.13) and $\widehat{V}_{gc}$ given by (4.4.17). Note that the combined estimating equation method converges to the mean estimating equation method as $g \rightarrow \infty$.

Table 4.7 reports the percent of negative variance estimates for the separate estimating equations and combined estimating equation methods. We observe from Table 4.7 that the separate estimating equations method produces negative variance estimates quite often even for g as large as 4000: proportion of negative variance estimates is 16% for the parameter β_1 . On the other hand, the combined estimating equation method performed well with 0% proportion of negative variance estimates, even for $g = 500$.

Table 4.6: Means and MSE estimators of β_0 and β_1 and coverage probabilities of associated 0.95-level normal theory intervals based on 1000 simulations: logistic linear regression with $\beta_0 = -1.0986$, $\beta_1 = 0.4796$ and $K = 100$ clusters.

	MEE	SEE	IEE
<i>Ignorable cluster sizes</i>			
E(est.)	-1.115	-1.126	-1.116
β_0 : MSE(est.) $\times 1000$	43.792	54.019	43.836
RB(var. est.) $\times 100$	-0.210	—	-0.310
Coverage probability	0.939	—	0.944
E(est.)	0.488	0.502	0.490
β_1 : MSE(est.) $\times 1000$	75.685	86.983	75.720
RB(var. est.) $\times 100$	8.080	—	5.950
Coverage probability	0.970	—	0.964
<i>Nonignorable cluster sizes</i>			
E(est.)	-1.113	-1.134	-1.510
β_0 : MSE(est.) $\times 1000$	59.856	59.042	214.637
RB(var. est.) $\times 100$	-6.633	—	-79.488
Coverage probability	0.942	—	0.510
E(est.)	0.480	0.500	0.499
β_1 : MSE(est.) $\times 1000$	106.563	106.490	79.356
RB(var. est.) $\times 100$	-5.510	—	0.329
Coverage probability	0.945	0.944	0.950

MEE, mean estimating equation; SEE, separate estimating equations; IEE, independence estimating equations; MSE, mean squared error; RB, relative bias.

Table 4.7: *Proportion of negative variance estimates(%) under the nested error model (4.8.38) with $K = 12$ clusters and n_i cluster sizes ranging from 1 to 5: 1000 simulation runs, $\beta_0 = 51$, $\beta_1 = 0.329$, $\beta_2 = -0.134$, $\sigma_\alpha^2 = 140$ and $\sigma_\epsilon^2 = 150$.*

	g=500		g=1000		g=4000	
	CEE	SEE	CEE	SEE	CEE	SEE
β_0	0	12.1	0	11.6	0	11.3
β_1	0	15.5	0	15.5	0	16.3
β_2	0	7.5	0	7.7	0	7.0

CEE, combined estimating equation; SEE, separate estimating equations.

Chapter 5

Analyses of Cluster-Correlated Binary Outcomes Data

5.1 Introduction

The traditional conditional logistic regression (CLR) is a commonly used method for analyzing cluster-correlated binary data from retrospective epidemiological, toxicological or other biological related studies. The mixed-effects logistic regression model is usually the working model adopted for this type of data. The assumptions underlying this model induce an exchangeable correlation structure among the binary responses within a cluster. The CLR method estimates the logistic regression parameters while treating the correlation structure arising from the random cluster effects as nuisance. However, for more complex correlation structures induced by additional unmeasured factors, the use of the CLR approach can lead to inefficient regression estimators and misleading statistical inferences. Rieger and Weinberg (2002) proposed

a simple but computationally intensive method that leads to asymptotically valid inferences whether the correlation structures are induced by random cluster effects or by additional unmeasured factors. In this method a pair of observations from each cluster - one from the case group and the other from the control group - is selected at random to obtain a sample of independent paired observations, assuming that each cluster contains at least one affected and one unaffected individual. This subsampling is repeated independently to obtain a large number, g , of subsamples of paired observations that are conditionally independent. Standard logistic regression method for matched-pair case-control designs can then be implemented separately on each of the g subsamples to obtain the regression estimators $\hat{\beta}_1^*, \dots, \hat{\beta}_g^*$ and the associated covariance estimators $\hat{V}_1^*, \dots, \hat{V}_g^*$. The resulting resampling estimator of the regression parameter is then given by $\hat{\beta}_{gs} = g^{-1} \sum_{j=1}^g \hat{\beta}_j^*$ with the associated covariance estimator given by

$$\hat{V}_{gs} = \frac{1}{g} \sum_{j=1}^g \hat{V}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\beta}_j - \hat{\beta}_{gs}) (\hat{\beta}_j - \hat{\beta}_{gs})'. \quad (5.1.1)$$

We propose a simple method, mean conditional estimating equation, that avoids resampling and yet leads to asymptotically valid inferences whether the within-cluster correlation structures are generated solely by the random effects or by additional unmeasured factors.

The rest of this chapter is organized as follows. In Section 5.2, the conditional logistic regression method is introduced. The method by Rieger and Weinberg (2002) is summarized in Section 5.3. The properties of the mean conditional estimating equation are discussed in Section 5.4. In Section 5.5, we present results of limited

simulation studies to assess the performance of the methods.

5.2 Conditional Logistic Regression

Suppose $s = \{(\mathbf{x}_{ij}, y_{ij}); j = 1, \dots, n_i; i = 1, \dots, K\}$ is a cluster-correlated data from a matched case-control study, where $y_{ij} \in \{0, 1\}$ is the case-control indicator of the j -th unit in the i -th cluster, $\mathbf{x}_{ij}$ is a $p \times 1$ vector of covariates associated with the j -unit in the i -th cluster, n_i is the number of units in the i -th cluster and K is the number of clusters. Note that $n_i = n_{0i} + n_{1i}$, where n_{0i} is the size of the control group in the i -th cluster and n_{1i} is the size of the case group in the i -th cluster. A mixed-effects logistic regression model for exploring the relationship between the response, the covariate and the nuisance parameters (cluster effects) is given by

$$\log \left\{ \frac{E(y_{ij} | \alpha_i, \mathbf{x}_{ij})}{1 - E(y_{ij} | \alpha_i, \mathbf{x}_{ij})} \right\} = \alpha_i + \mathbf{x}_{ij}' \boldsymbol{\beta}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, K, \quad (5.2.2)$$

where α_i is the i -th cluster effect and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters. We note that for the model (5.2.2), the regression parameter measures the change in the conditional logit of the probability of response with the covariate $\mathbf{x}$ for individuals in each of the underlying risk groups described by α_i .

There are two standard ways of analyzing the mixed-effects model (5.2.2): First, one can assume that the unobserved cluster effects, $\alpha_1, \dots, \alpha_K$, represent independent identically distributed sample from a parametric family of distributions with density f_θ (where θ is assumed unknown). Estimators of $\boldsymbol{\beta}$ and θ are then obtained by jointly maximizing an appropriate likelihood with respect to $\boldsymbol{\beta}$ and θ , see Neuhaus and Jewell (1990).

The second method for analyzing the mixed-effects logistic regression model is the conditional logistic regression approach. This method makes no distributional assumptions on the cluster effects $\alpha_1, \dots, \alpha_K$, rather it eliminates the nuisance parameters α_i by conditioning on their sufficient statistics. Hereafter, our discussion will focus on this second method of analysis.

We first note that the cluster-specific likelihood function of β is given by

$$L_i = L(\beta | \alpha_i, \mathbf{y}_i, \mathbf{x}_i) = \frac{\exp \left(\alpha_i \sum_{j=1}^{n_i} y_{ij} + \sum_{j=1}^{n_i} y_{ij} \mathbf{x}'_{ij} \beta \right)}{\prod_{j=1}^{n_i} \left\{ 1 + \exp(\alpha_i + \mathbf{x}'_{ij} \beta) \right\}}. \quad (5.2.3)$$

The sufficient statistics for α_i and β are $S_i = \sum_{j=1}^{n_i} Y_{ij}$ and $\mathbf{T} = \sum_{i=1}^K \mathbf{T}_i$ respectively, where $\mathbf{T}_i = \sum_{j=1}^{n_i} Y_{ij} \mathbf{X}_{ij}$ is the contribution of the i -th cluster to $\mathbf{T}$. In a matched case-control study, the set of sufficient statistics, $\{S_i; i = 1, \dots, K\}$, is fixed in advance. Let Ω_i denote the set of possible values of $\mathbf{T}_i$ such that $S_i = s_i$ and $c(\mathbf{t}_i)$ represent the number of distinct ordered sets of $\{y_{ij}; j = 1, \dots, n_i\}$ giving rise to $\mathbf{T}_i = \mathbf{t}_i$ subject to $S_i = s_i$. Then the conditional likelihood of $\mathbf{T}_i = \mathbf{t}_i$ given $S_i = s_i$ is

$$L = \prod_{i=1}^K P(\mathbf{T}_i = \mathbf{t}_i | S_i = s_i) = \prod_{i=1}^K \left\{ \frac{c(\mathbf{t}_i) \exp(\mathbf{t}'_i \beta)}{\sum_{\mathbf{a} \in \Omega_i} c(\mathbf{a}) \exp(\mathbf{a}' \beta)} \right\}. \quad (5.2.4)$$

We note that there are $\prod_{i=1}^K C_{s_i}^{n_i}$ possible sequences $\{y_{ij}\}$ for each $\mathbf{t}$ obtained, $C_{s_i}^{n_i}$ represents a permutation of selecting s_i from n_i .

The conditional logistic regression method estimates the regression parameter β by solving the conditional maximum likelihood score equation

$$U_c(\beta) = \sum_{i=1}^K \left\{ \mathbf{t}_i - \frac{\sum_{\mathbf{a} \in \Omega_i} \mathbf{a} c(\mathbf{a}) \exp(\mathbf{a}' \beta)}{\sum_{\mathbf{a} \in \Omega_i} c(\mathbf{a}) \exp(\mathbf{a}' \beta)} \right\} = \sum_{i=1}^K \mathbf{u}_i(\beta) = \mathbf{0}. \quad (5.2.5)$$

In general (5.2.5) is solved iteratively to obtain the solution $\hat{\beta}_c$. The Newton-Raphson

algorithm is given by

$$\hat{\beta}^r = \hat{\beta}^{(r-1)} + \{J_c(\hat{\beta}^{(r-1)})\}^{-1} U_c(\hat{\beta}^{(r-1)}), \quad (5.2.6)$$

where $\hat{\beta}^{(r-1)}$ is the value of β obtained at the $(r-1)$ -th iteration, and $U_c(\hat{\beta}^{(r-1)})$ and $J_c(\hat{\beta}^{(r-1)})$ are the values of $U(\beta)$ and $J(\beta) = -\partial U(\beta)/\partial \beta'$ evaluated at $\beta = \hat{\beta}^{(r-1)}$.

Iterating the algorithm to convergence produces the estimator $\hat{\beta}_c$ as well as observed information matrix $J_c(\hat{\beta}_c)$. An asymptotic variance estimator of $\hat{\beta}_c$ is

$$\widehat{V}(\hat{\beta}_c) = \left\{ \sum_{i=1}^K \frac{\partial u_i(\hat{\beta}_c)}{\partial \beta'} \right\}^{-1}. \quad (5.2.7)$$

Implementation of the CLR method in general requires specialized computer software such as STATA. However, for data from matched case-control studies, one can modify existing syntax in some of the standard softwares to implement the CLR method. Appendix II shows a syntax in Splus.

5.2.1 Matched-Pair Case-Control Design

When each cluster is of size two the conditional likelihood (5.2.4) reduces to

$$L_{\Delta}(\beta) = \prod_{i=1}^K P(T_i = x_{i1} | S_i = 1) = \prod_{i=1}^K \frac{\exp(x'_{\Delta i} \beta)}{1 + \exp(x'_{\Delta i} \beta)}, \quad (5.2.8)$$

where $x_{\Delta i} = x_{i1} - x_{i0}$, x_{i1} and x_{i0} are the i -th covariates from the case and control groups respectively. This means that for this matched-pair case-control design, standard logistic regression software can be implemented to estimate the regression parameters and their standard errors. To do this, $x_{\Delta i}$ is set as the covariate, $y_i = 1$ and the sample is of size K . Note that all terms involving cluster-specific covariates

are eliminated as well as the nuisance parameters α_i , see Hosmer and Lemeshow (1989 page 190) and Rieger and Weinberg (2002).

As noted by Rieger and Weinberg (2002), when all unmeasured cluster-specific factors cannot be aggregated into a cluster-specific baseline risk or when a more complex correlation structure is associated with the data, then the CLR approach can lead to invalid inferences. In the next section, we summarize the method proposed by Rieger and Weinberg (2002) as an alternative approach to the CLR but also remains valid when the assumptions for CLR are violated.

5.3 Separate Conditional Estimating Equations

In this section, we summarize the separate conditional estimating equations (SCEE) method proposed by Rieger and Weinberg (2002). Suppose that within each cluster there is at least one unit in the case group and one unit in the control group. Two units are selected at random from each cluster - one from the case group and the other from the control group - resulting in a subsample, s^* , of independent matched-pair observations of size K , where K is the number of clusters. The procedure is repeated a large number of times, say g , leading to g subsamples, s_j^* ($j = 1, \dots, g$) that are conditionally independent given the original sample. We note that each subsample is similar to a sample from a matched-pair case-control design discussed in Subsection 5.2.1. Hence, the method of analysis discussed for the matched-pair case-control design can be adopted to analyze each subsample. The subsampling estimators, $\hat{\beta}_j^*$, $j = 1, \dots, g$ are obtained by solving g separate conditional estimating

equations;

$$U_j^*(\beta) = \sum_{i=1}^K u_{ij}(\beta) = 0; \quad j = 1, \dots, g, \quad (5.3.9)$$

where $u_{ij} = x_{\Delta ij} - x_{\Delta ij} \exp(x'_{\Delta ij} \beta) / \{1 + \exp(x'_{\Delta ij} \beta)\}$, $x_{\Delta ij} = x_{ij1} - x_{ij0}$, x_{ij1} and x_{ij0} are the $p \times 1$ covariates of the j -th unit of the i -th cluster from case group and control group respectively. In general, we require g iterative solutions to get $\hat{\beta}_1^*, \dots, \hat{\beta}_g^*$ using (5.3.9). Standard logistic regression software may be used repeatedly to obtain the subsampling estimators, $\hat{\beta}_j^*, j = 1, \dots, g$ and the associated covariance estimators, $\widehat{V}_1^*, \dots, \widehat{V}_g^*$. The separate conditional estimating equations estimator of β is given by

$$\hat{\beta}_{gS} = \frac{1}{g} \sum_{j=1}^g \hat{\beta}_j^*. \quad (5.3.10)$$

The repeated subsampling covariance estimator of $\hat{\beta}_{gS}$ is given by

$$\widehat{V}_{gS} = \frac{1}{g} \sum_{j=1}^g \widehat{V}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\beta}_j^* - \hat{\beta}_{gS})(\hat{\beta}_j^* - \hat{\beta}_{gS})', \quad (5.3.11)$$

where

$$\widehat{V}_j^* = \left\{ \sum_{i=1}^K \frac{\partial u_{ij}(\hat{\beta}_j^*)}{\partial \hat{\beta}'} \right\}^{-1}. \quad (5.3.12)$$

The variance estimators of the components of $\hat{\beta}_{gS}$ can take negative values if the number of clusters, K , is small.

5.4 Mean Conditional Estimating Equation

Instead of solving g separate conditional estimating equations (5.3.9), we combine these equations before solving for β :

$$U_{gc}(\beta) = \frac{1}{g} \sum_{j=1}^g U_j^*(\beta) = \frac{1}{g} \sum_{j=1}^g \sum_{i=1}^K u_{ij}(\beta) = 0. \quad (5.4.13)$$

In general, we solve (5.4.13) using the Newton-Raphson iterations (5.2.6) with $U_c(\hat{\beta}^{(r-1)})$ changed to $U_{gc}(\hat{\beta}^{(r-1)})$ and $J_c(\hat{\beta}^{(r-1)})$ changed to $J_{gc}(\hat{\beta}^{(r-1)})$, where

$$J_{gc}(\beta) = -\frac{\partial U_{gc}(\beta)}{\partial \beta'} = \frac{1}{g} \sum_{j=1}^g \sum_{i=1}^K \frac{\partial u_{ij}(\beta)}{\partial \beta'}. \quad (5.4.14)$$

At convergence, we get the estimator $\hat{\beta}_{gc}$ as well as the observed information matrix $J_{gc}(\hat{\beta}_{gc})$. The covariance estimator of $\hat{\beta}_{gc}$ is given by

$$\widehat{V}(\hat{\beta}_{gc}) = \{J_{gc}(\hat{\beta}_{gc})\}^{-1} \left[\frac{1}{g} \sum_{j=1}^g \{u_{ij}(\hat{\beta}_{gc})u_{ij}(\hat{\beta}_{gc})' - U_j^*(\hat{\beta}_{gc})U_j^*(\hat{\beta}_{gc})'\} \right] \times \{J_{gc}(\hat{\beta}_{gc})\}^{-1}. \quad (5.4.15)$$

Some basic properties of the combined conditional estimating function are given in Theorem 5.1.

Theorem 5.1 1. Conditional on the original sample s , $U_{gc}(\beta)$ converges to

$E\{U_{gc}(\beta)|s\} = U_{\infty}(\beta)$ as $g \rightarrow \infty$, where

$$U_{\infty}(\beta) = \sum_{i=1}^K \frac{1}{n_{i1}n_{i0}} \sum_{j=1}^{n_{i1}} \sum_{r=1}^{n_{i0}} u_{ijr}(\beta), \quad (5.4.16)$$

$$u_{ijr}(\beta) = (x_{ij1} - x_{ir0}) - \frac{(x_{ij1} - x_{ir0}) \exp\{(x_{ij1} - x_{ir0})'\beta\}}{[1 + \exp\{(x_{ij1} - x_{ir0})'\beta\}]},$$

x_{ij1} is a covariate from the case group, x_{ij0} a covariate from the control group, n_{i0} is the number of units in the control group from the i -th cluster and n_{i1} is the number of units in the case group from the i -th cluster. Note that $n_i = n_{i0} + n_{i1}$.

2. $E\{U_{gc}(\beta)\} = E\{U_1^*(\beta)\}.$
3. $var\{U_{gc}(\beta)\} = var\{U_\infty(\beta)\} + \frac{1}{g}E[var\{U_1^*(\beta)|s\}].$
4. Let $\hat{\beta}_\infty$ be a solution to $U_\infty(\beta) = 0$ and $\hat{\beta}_{gc}$ be a solution to $U_{gc}(\beta) = 0$. Then under mild regularity conditions, $\hat{\beta}_{gc}$ and $\hat{\beta}_\infty$ are asymptotically equivalent as $g \rightarrow \infty$ and $K \rightarrow \infty$.

Proof

1. Proof of Result 1 follows by noting that conditional on the original sample s , $U_1^*, \dots, U_g^*$ are independent identically distributed bounded random vectors.
2. Result 2 follows from the standard relationship between conditional and unconditional expectations:

$$E\{U_{gc}(\beta)\} = E[E\{U_1^*(\beta)|s\}] = E\{U_1^*(\beta)\}.$$

3. This follows from the corresponding results for variances, and the conditional independence of the U_j^* 's given the original sample s :

$$\begin{aligned} var\{U_{gc}(\beta)\} &= var[E\{U_{gc}(\beta)|s\}] + E[var\{U_{gc}(\beta)|s\}] \\ &= var\{U_\infty(\beta)\} + \frac{1}{g}E[var\{U_1^*(\beta)|s\}]. \end{aligned}$$

4. Consider the following regularity conditions:

C1. For some $\delta > 0$,

$$\sup_i E \left\| \frac{\partial u_{ijr}(\beta)}{\partial \beta'} \right\|^{1+\delta} < \infty$$

C2. Let

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{i=1}^K E \left\{ -\frac{\partial}{\partial \beta'} \frac{1}{n_{i1} n_{i0}} \sum_{j=1}^{n_{i1}} \sum_{r=1}^{n_{i0}} u_{ijr}(\beta) \right\} = \lim_{K \rightarrow \infty} \frac{1}{K} \sum_{i=1}^K E \left\{ -\frac{\partial u_i(\beta)}{\partial \beta'} \right\} = I(\beta),$$

where $I(\beta)$ exists and it is nonsingular.

C3. For some $\delta > 0$

$$\sup_i E \|u_{ijr}(\beta)\|^{2+\delta} < \infty.$$

C4. Let

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{i=1}^K \text{Var} \{u_i(\beta)\} = V_U,$$

a positive-definite covariance matrix. By conditions C1, C2 and the Taylor's expansion of $U_\infty(\hat{\beta}_\infty)$ around β we have,

$$\hat{\beta}_\infty - \beta = \{I(\beta)\}^{-1} \frac{1}{K} U_\infty(\beta) + o_p(1/\sqrt{K}). \quad (5.4.17)$$

Similarly,

$$\hat{\beta}_{gc} - \beta = \{I(\beta)\}^{-1} \frac{1}{K} U_{gc}(\beta) + o_p(1/\sqrt{K}). \quad (5.4.18)$$

From (5.4.17) and (5.4.18) we have

$$\|\hat{\beta}_\infty - \hat{\beta}_{gc}\| \leq \|\{I(\beta)\}^{-1}\| \frac{1}{K} \|U_\infty(\beta) - U_{gc}(\beta)\| + o_p(1/\sqrt{K}) \quad (5.4.19)$$

Apply the law of large numbers to the resampling distribution, the first term converges to zero in probability as $g \rightarrow \infty$. Hence, as $g \rightarrow \infty$ we get $\hat{\beta}_\infty - \hat{\beta}_{gc} = o_p(1/\sqrt{K})$. Thus $\hat{\beta}_\infty$ and $\hat{\beta}_{gc}$ are asymptotically equivalent as $K \rightarrow \infty$.

■

From Result 1 of Theorem 5.1, it follows that instead of using the combined conditional estimating function, $U_{gc}(\beta)$, we can avoid repeated resampling by using the mean conditional estimating function $U_\infty(\beta) = U_M(\beta)$ given by (5.4.16). Therefore, we focus on $U_M(\beta)$. The mean conditional estimating equation (MCEE) is given by

$$U_\infty(\beta) = \sum_{i=1}^K \frac{1}{n_{i1}n_{i0}} \sum_{j=1}^{n_{i1}} \sum_{r=1}^{n_{i0}} \mathbf{u}_{ijr}(\beta) = 0. \quad (5.4.20)$$

In general, we solve (5.4.20) using the Newton-Raphson iterations (5.2.6) with $U_c(\hat{\beta}^{(r-1)})$ changed to $U_M(\hat{\beta}^{(r-1)})$ and $J_c(\hat{\beta}^{(r-1)})$ changed to $J_M(\hat{\beta}^{(r-1)})$, where $U_M(\hat{\beta}^{(r-1)})$ and $J_M(\hat{\beta}^{(r-1)})$ are the values of $U_M(\beta)$ and $J_M(\beta)$ evaluated at $\beta = \hat{\beta}^{(r-1)}$. The covariance of $\hat{\beta}_M$ is given by

$$\widehat{V}(\hat{\beta}_M) = \{J_M(\hat{\beta}_M)\}^{-1} \left[\sum_{i=1}^K \left\{ \frac{1}{n_{i1}n_{i0}} \sum_{j=1}^{n_{i1}} \sum_{r=1}^{n_{i0}} \mathbf{u}_{ijr}(\hat{\beta}_M) \right\} \left\{ \frac{1}{n_{i1}n_{i0}} \sum_{j=1}^{n_{i1}} \sum_{r=1}^{n_{i0}} \mathbf{u}_{ijr}(\hat{\beta}_M) \right\}' \right] \times \{J_M(\hat{\beta}_M)\}^{-1} \quad (5.4.21)$$

Theorem 5.2 gives the asymptotic distribution of $\hat{\beta}_M$.

Theorem 5.2 *Under mild regularity conditions, and given that $\hat{\beta}_M$ is a consistent estimator of β , $K^{1/2}(\hat{\beta}_M - \beta)$ is asymptotically multivariate normal with zero mean and covariance matrix estimated by (5.4.21).*

Proof

By assumptions C3 and C4, and for any vector $\mathbf{a}$,

$$\begin{aligned} \frac{\sum_{i=1}^K E |\mathbf{a}' \mathbf{u}_i(\beta)|^{2+\delta}}{\left(\sum_{i=1}^K \text{Var}(\mathbf{a}' \mathbf{u}_i(\beta)) \right)^{\frac{2+\delta}{2}}} &= O(1/K^{\frac{\delta}{2}}). \\ &\rightarrow 0 \text{ as } K \rightarrow \infty \end{aligned}$$

Hence by Lyapunov's central limit theorem

$$\frac{1}{\sqrt{K}}U_M(\beta) = \frac{1}{\sqrt{K}}\sum_{i=1}^K \mathbf{u}_i(\beta) \xrightarrow{D} N(0, V_U). \quad (5.4.22)$$

Using Taylor expansion of $U_M(\hat{\beta}_M)$ around β , assumption C2 and (5.4.22) we have

$$\begin{aligned} \sqrt{K}(\hat{\beta}_M - \beta) &= \left\{ -\frac{1}{K} \sum_{i=1}^K \frac{\partial \mathbf{u}_i(\beta)}{\partial \beta'} \Big|_{\beta=\tilde{\beta}} \right\}^{-1} \left\{ \frac{1}{\sqrt{K}} \sum_{i=1}^K \mathbf{u}_i(\beta) \right\} \\ &\xrightarrow{D} N(0, I(\beta)^{-1} V_U I(\beta)^{-1}), \end{aligned}$$

where $\tilde{\beta} \in (\hat{\beta}_M, \beta)$, noting that the covariance is consistently estimated by $\widehat{V}_M/K$, where $\widehat{V}_M$ is given by (5.4.21). ■

5.5 Simulation Study

We conducted a limited simulation study to compare the performance of the proposed method. The simulation data were generated following a set-up by Rieger and Weinberg (2002) for a mixed-effects model with common cluster effect. In this case as expected, both the CLR and the MCEE methods lead to asymptotically valid inferences. Consider a logistic regression model

$$\log \left\{ \frac{E(y_{ij}|x_{ij})}{1 - E(y_{ij}|x_{ij})} \right\} = \alpha + x_{ij}\beta; \quad j = 1, \dots, n_i; i = 1, \dots, K, \quad (5.5.23)$$

where x_{ij} is an explanatory variable attached to the j -th unit in the i -th cluster, y_{ij} is the corresponding binary outcome variable, n_i is the i -th cluster size, K is the

number of clusters, α is a common cluster effect and β is the regression parameter. The regression parameter was set at $\beta = 0.3$, x_{ij} takes the value 1 if $U_{x_{ij}} \leq 1/2$ or the value 0 otherwise, where $U_{x_{ij}}$ is from a uniform distribution, $[0, 1]$. The cluster sizes were generated from a truncated binomial distribution $(8, 1/2)$, discarding cluster sizes of zero or 1. The common cluster effect was fixed at -1.0986 (in other words a baseline risk of 0.25). The response variable takes the value 1 if $U_{y_{ij}} < E(y_{ij}|x_{ij})$ or the value 0 otherwise, where $U_{y_{ij}}$ is from a uniform distribution, $[0, 1]$. This set-up leads to intra-cluster correlation coefficient, $\rho = 0$. For $K = 100$, 1000 distinct samples were generated.

Table 5.1 reports the following values (i) the mean of the estimator $\hat{\beta}$, given by $\hat{\beta}(\cdot) = \sum_{r=1}^R \hat{\beta}(r)/R$; (ii) mean squared error of $\hat{\beta}$, given by $MSE(\hat{\beta}) = \sum_{r=1}^R \{\hat{\beta}(r) - \beta\}^2/R$; (iii) relative bias of a variance estimator $v(\hat{\beta})$, given by $RB\{v(\hat{\beta})\} = [E\{v(\hat{\beta})\} - MSE(\hat{\beta})]/MSE(\hat{\beta})$, where $E\{v(\hat{\beta})\} = \sum_{r=1}^R v_r(\hat{\beta})/R$ and $v_r(\hat{\beta})$ denotes the variance estimator obtained from the r -th simulated sample; (iv) coverage probability associated with 95% normal theory confidence interval on $\beta : \{\hat{\beta} - 2v(\hat{\beta})^{1/2}, \hat{\beta} + 2v(\hat{\beta})^{1/2}\}$. The coverage probability in (iv) was calculated as the proportion of the simulated intervals, $\hat{\beta}_r \pm v_r(\hat{\beta})^{1/2}$, that cover the true β . Table 5.1 shows that both methods perform well with respect to bias of the estimator. The estimators $\hat{\beta}_c$ and $\hat{\beta}_M$ are also similar with respect to the mean squared error. The relative bias of the variance estimator is less than 10% for both methods. Furthermore, both methods perform well with respect to the coverage probabilities of 0.95-level normal intervals.

In this chapter we have demonstrated that the mean conditional estimating equation method is an asymptotically valid method for analyzing cluster-correlated binary

Table 5.1: Means and mean squared errors of estimators of β , relative biases of variance estimators and coverage probabilities of associated 0.95 normal theory intervals based on 1000 simulations; $\beta = 0.3$ and $K = 100$ clusters

	CLR	MCEE
$E(\text{est.})$	0.3058	0.3047
$MSE(\text{est.})$	0.0492	0.0528
$RB(\text{var. est.})$	-0.0758	-0.0620
Coverage probability	0.946	0.944

CLR, Conditional logistic regression; MCEE, mean conditional estimating equation; MSE = mean squared error; RB = relative bias.

data outcomes. We also alluded to the fact that when the correlation structure associated with the data are generated by unmeasured covariate factors other than or in combination with random cluster effects, the CLR method may lead to asymptotically invalid inferences. However, MCEE remains valid whether the correlation structure is generated by mixed-effects model or by unmeasured covariate factors or by more complex random factors. In addition, since the MCEE method is not a resampling method, it is computationally simpler than the SCEE method. Also, for small number of clusters, the SCEE method can lead to negative variance estimators whereas the MCEE lead to positive definite variance estimates.

It is important to note that further simulation studies are needed to assess the performance (empirically) of the proposed methods under conditions where the CLR

method fails. We propose to explore this in future research work.

Chapter 6

Analysis of Data With Nonignorable Cluster sizes and Two Types of Correlation

6.1 Introduction

Many methods for the analysis of cluster-correlated data are well known. The model formulations of most of these methods, including the very popular Liang and Zeger (1986) generalized estimating equations (GEE), account for the dependence between observations from the same cluster by either directly modeling the correlation or by assuming a working correlation structure.

It is well known that ignoring the intraclass correlation decreases the standard error of the regression parameter estimators, hence overestimating the statistical significance of the associated covariates. Most of these methods, however, are suitable

only for analyzing data with a single direction of clustering.

One problem which has received less attention is the analysis of data with more than one direction of clustering. In this chapter, we propose an inverse-sampling combined estimating equation approach for analyzing data with two types of correlations: longitudinal correlation and within-cluster correlation.

An example of a study where the data may have two directions of clustering was a study from the Waterloo Smoking Prevention Project 3 (WSPP3). This project was a third in the series of randomized controlled smoking prevention trials, designed to develop, evaluate and disseminate an effective school-based social-influences smoking prevention program (Best et al., 1995 and Brown & Cameron, 1997). In each of the three series of the study, data were collected on students from selected schools. Since students from the same school tend to be similar, the observations collected on these students for any one series may be correlated (this type of correlation is sometimes referred to as familial correlation). The repeated observations collected on students from series to series may give rise to another type of correlation, namely longitudinal correlation. The resulting data from the three series of the study may therefore have more than one type of correlation (two to be exact): correlation among students from the same school and correlation as a result of repeated observations from the same student over the three series.

As in the case of cluster correlated data discussed in Chapter 4, methods for the analysis of data with two types of correlations that do not account for these correlations may tend to decrease the standard errors of the regression parameter estimators. This may lead to overestimation of the statistical significance of the

covariates associated with these regression parameters.

Two primary approaches used to account for the correlation in estimating the regression parameters are either direct modeling of the correlation or by assuming a working correlation matrix. The latter approach is the most commonly used when inference on the regression parameters is the primary interest of investigation. When analyzing longitudinal data, it is well known that the regression parameter estimators based on independence estimating equations method (Liang and Zeger, 1986) are consistent but may be inefficient, see Crowder (1995). However, for most practical purposes, the loss in efficiency may be negligible, see Sutradhar and Das (1999).

For within-cluster and longitudinally correlated data, we study the efficiency of the regression parameter estimators based on a working correlation matrix. Results of the study are presented in Section 6.2.

Most of the methods for analyzing cluster-correlated data with repeated observations implicitly assume that the cluster sizes are ignorable. When this assumption fails, however, the resulting inferences may be asymptotically invalid. As in the case of the analysis of cluster-correlated data discussed extensively in Chapter 4, we proposed three methods that remain asymptotically valid whether the cluster sizes are informative or noninformative. The methods are: (i) separate longitudinal independence estimating equations (SLIEE), (ii) combined longitudinal independence estimating equation (CLIEE) and (iii) mean longitudinal independence estimating equation (MLIEE). Suppose there are K clusters each with n_i ($i = 1, \dots, K$) units and the i -th unit has T_{ij} ($j = 1, \dots, n_i$) repeated observations. For the SLIEE and the CLIEE methods, a unit is selected at random from each cluster to obtain K inde-

pendent units each of size T_{ij} . The procedure is repeated a large number, say g times. The SLIEE method analyzes each of the g inverse-samples separately using standard longitudinal independence estimating equations (IEE) method. The SLIEE regression parameter estimator is obtained by averaging. The CLIEE method combines all the g separate equations from SLIEE before solving for the regression parameter. On the other hand, the MLIEE method avoids resampling, and it is motivated by the CLIEE.

This chapter is organized as follows: In Section 6.2 we introduce a marginal model to analyze data with two types of correlations. We then use this model to study the efficiency of the regression parameter estimators based on assumed working correlation structures and an implicit assumption that the cluster sizes are ignorable. Also as alluded to earlier, methods for analyzing correlated data that implicitly assume that cluster sizes are ignorable can lead to misleading results when this assumption fails. In Section 6.3, an illustrative example using a total independence estimating equations (TIEE) is discussed. A nested error linear regression model is used to illustrate the bias of the TIEE estimators under nonignorable cluster sizes. In Section 6.4 we introduce SLIEE, CLIEE and MLIEE methods. We also present basic properties of the proposed methods.

6.2 A Marginal Model

Suppose x_{ijt} ($t = 1, \dots, T_{ij}$; $j = 1, \dots, n_i$; $i = 1, \dots, K$) represents a $p \times 1$ vector of covariates corresponding to a response variable y_{ijt} for the t -th outcome of the j -th

unit from the i -th cluster, where T_{ij} is the number of repeated observations of the j -th unit, n_i is the number of units in the i -th cluster and K is the number of clusters. Let $s = \{(\mathbf{x}_{ijt}, y_{ijt}); t = 1, \dots, T_{ij}; j = 1, \dots, n_i; i = 1, \dots, K\}$ represent the sample of observations and the cluster sizes, n_i , may be ignorable or nonignorable. For the moment we assume that the cluster sizes are ignorable. Assume the regression model of interest is given by

$$\mu_{ijt} = E(y_{ijt}) = \mu(\mathbf{x}'_{ijt}\boldsymbol{\beta}) = \mu(\eta_{ijt}), \quad (6.2.1)$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression parameters. Let $\boldsymbol{\Sigma}_i(\phi, \alpha) = \mathbf{A}_i^{\frac{1}{2}} \mathbf{C}_i(\phi, \alpha) \mathbf{A}_i^{\frac{1}{2}}$ be the $T_i \times T_i$ covariance matrix of $\mathbf{y}_i$, the T_i - vector of responses from the i -th cluster, where $T_i = \sum_{j=1}^{n_i} T_{ij}$, $\mathbf{A}_i = \text{diag}\{\text{var}(y_{ijt})\}$, $\mathbf{C}_i(\phi, \alpha)$ is the $T_i \times T_i$ correlation matrix characterized by the within-cluster correlation parameters, ϕ , and the longitudinal correlation parameters, α .

The quasilielihood estimator, $\hat{\boldsymbol{\beta}}_Q$, of $\boldsymbol{\beta}$ is the solution to the score equation given by

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^K \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' \boldsymbol{\Sigma}_i^{-1}(\phi, \alpha) (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}, \quad (6.2.2)$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{ij}, \dots, y_{in_i})'$, $\mathbf{y}_{ij} = (y_{ij1}, \dots, y_{ijt}, \dots, y_{ijT_{ij}})'$, $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = \boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{ij}, \dots, \mu_{in_i})'$ and $\boldsymbol{\mu}_{ij} = (\mu_{ij1}, \dots, \mu_{ijt}, \dots, \mu_{ijT_{ij}})'$. In general, the true correlation matrix $\mathbf{C}_i(\phi, \alpha)$ is unknown, and hence to solve (6.2.2) for $\boldsymbol{\beta}$, we replace the correlation matrix by a 'working' correlation matrix.

The Newton-Raphson (N-R) algorithm is commonly used to solve (6.2.2). The r -th step of the N-R algorithm is given by

$$\hat{\boldsymbol{\beta}}^r = \hat{\boldsymbol{\beta}}^{(r-1)} + \{\mathbf{J}(\hat{\boldsymbol{\beta}}^{(r-1)})\}^{-1} \mathbf{U}(\hat{\boldsymbol{\beta}}^{(r-1)}), \quad (6.2.3)$$

where $\hat{\beta}^{(r-1)}$ is the value of $\hat{\beta}$ obtained at the $(r-1)$ -th iteration, and $U(\hat{\beta}^{(r-1)})$ and $J(\hat{\beta}^{(r-1)})$ are the values of $U(\beta)$ and $J(\beta) = -\partial U(\beta)/\partial \beta'$ evaluated at $\beta = \hat{\beta}^{(r-1)}$.

Iterating the algorithm to convergence produces the estimator $\hat{\beta}_Q$ as well as the observed information matrix $J(\hat{\beta}_Q)$. A linearization covariance matrix of $\hat{\beta}_Q$ is given by

$$V_Q = V(\hat{\beta}_Q) = \left[\sum_{i=1}^K \left(\frac{\partial \mu_i}{\partial \beta} \right)' \left\{ A_i^{\frac{1}{2}} C_i(\phi, \alpha) A_i^{\frac{1}{2}} \right\}^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right) \right]^{-1}. \quad (6.2.4)$$

6.2.1 Relative Efficiencies

In this subsection, we study the efficiency of the regression parameter estimators $\hat{\beta}_W$, of β based on assumed working correlation structures. A simple approach we have adopted for studying the effect of assuming working correlation structures on the efficiency of the regression parameter estimators is through the computation of asymptotic relative efficiencies. For the purposes of these computations, we let $T_{ij} = T$. We also assume that the correlation matrix of $\mathbf{y}_i$ is of the form $C_i(\phi, \alpha) = C_{1i}(\phi) \otimes C_{2i}(\alpha)$, where $\otimes$ is the Kronecker product, $C_i(\phi)$ is an $n_i \times n_i$ matrix of within-cluster correlations and $C_{2i}(\alpha)$ is a $T \times T$ matrix of longitudinal correlations.

Let $V_i = A_i^{\frac{1}{2}} C_i^*(\phi, \alpha) A_i^{\frac{1}{2}}$ represent a working covariance matrix, where $C_i^*(\phi, \alpha)$ is as defined in (i), (ii), (iii) and (iv) below. We consider the following working correlation matrices: (i) within-cluster correlation is ignored (within-cluster independence): $C_i^*(\phi, \alpha) = I_{n_i} \otimes C_{2i}(\alpha)$; (ii) longitudinal correlation is ignored (longitudinal independence): $C_i^*(\phi, \alpha) = C_{1i}(\phi) \otimes I_T$; (iii) both within-cluster and longitudinal correlations are ignored (total independence): $C_i^*(\phi, \alpha) = I_{n_i}(\phi) \otimes I_T$, where I_{n_i} is an

$n_i \times n_i$ identity matrix, and I_T is a $T \times T$ identity matrix; (iv) a working longitudinal correlation matrix: $C_i^*(\phi, \alpha) = C_{1i}(\phi) \otimes C_{2i}^*(\alpha)$, where $C_{2i}^*(\alpha) \neq C_{2i}(\alpha)$.

Under the working covariance assumption, $V_i = A_i^{\frac{1}{2}} C_i^*(\phi, \alpha) A_i^{\frac{1}{2}}$, a consistent estimator, $\hat{\beta}_W$, of β is the solution to the estimating equations:

$$U_W(\beta) = \sum_{i=1}^K \left(\frac{\partial \mu_i}{\partial \beta} \right)' \widehat{V}_i^{-1} (y_i - \mu_i(\beta)) = 0, \quad (6.2.5)$$

where $\widehat{V}_i = V_i(\hat{\phi}, \hat{\alpha})$, $\hat{\alpha}$ and $\hat{\phi}$ are estimators of α and ϕ respectively such that $\hat{\alpha} \rightarrow_p \alpha$ and $\hat{\phi} \rightarrow_p \phi$ for cases (i), (ii) and (iii), $\hat{\alpha} \rightarrow_p \alpha_0 \neq \alpha$ for case (iv).

A linearization covariance matrix of $\hat{\beta}_W$ is given by

$$V_W = V(\hat{\beta}_W) = F_W^{-1} F F_W^{-1}, \quad (6.2.6)$$

where

$$F_W = \sum_{i=1}^K \left(\frac{\partial \mu_i}{\partial \beta} \right)' \left\{ A_i^{\frac{1}{2}} C_i^*(\phi, \alpha) A_i^{\frac{1}{2}} \right\}^{-1} \left(\frac{\partial \mu_i}{\partial \beta} \right), \quad (6.2.7)$$

and

$$F = \sum_{i=1}^K \left(\frac{\partial \mu_i}{\partial \beta} \right)' \left\{ A_i^{-\frac{1}{2}} C_i^*(\phi, \alpha)^{-1} C_i(\phi, \alpha) C_i^*(\phi, \alpha)^{-1} A_i^{-\frac{1}{2}} \right\} \left(\frac{\partial \mu_i}{\partial \beta} \right). \quad (6.2.8)$$

We note that the linearization covariance matrix of $\hat{\beta}_W$ given in (6.2.6) reduces to the linearization covariance matrix of $\hat{\beta}_Q$ given in (6.2.4) if $C_i^*(\phi, \alpha) = C_i(\phi, \alpha)$

A measure of relative asymptotic efficiency of $\hat{\beta}_W$ to $\hat{\beta}_Q$ expressed as a percentage is given by

$$\text{Rel. Eff.}(\hat{\beta}_{Ql}, \hat{\beta}_{Wl}) = \frac{V(\hat{\beta}_{Ql})}{V(\hat{\beta}_{Wl})} \times 100, \quad l = 1, \dots, p \quad (6.2.9)$$

where $V(\hat{\beta}_{Wl})$ and $V(\hat{\beta}_{Ql})$ are the l -th diagonal elements of $V(\hat{\beta}_W)$ and $V(\hat{\beta}_Q)$ respectively.

For the relative asymptotic efficiency calculations, we considered the case of binary responses y_{ijt} and used the logistic regression model given by

$$\mu(x'_{ijt}\beta) = \frac{\exp(x'_{ijt}\beta)}{1 + \exp(x'_{ijt}\beta)}, \quad t = 1, \dots, T; \quad j = 1, \dots, n_i; \quad i = 1, \dots, K$$

where $\beta = (\beta_1, \beta_2)'$ and $\beta_1 = \beta_2 = 1$. The number of clusters $K = 100$, the number of units in each cluster $n_i = n = 4$, and the number of repeated observations $T = 4$. The first covariate was chosen as follows: $x_{1i1t} = x_{1i2t} = -1$, $x_{1i3t} = x_{1i4t} = 0$, ($t = 1, \dots, 4$), $i = 1, \dots, K/4$; $x_{1i1t} = x_{1i2t} = -1$, $x_{1i3t} = x_{1i4t} = 1$, ($t = 1, \dots, 4$), $i = K/4 + 1, \dots, K/2$; $x_{1i1t} = x_{1i2t} = 0$, $x_{1i3t} = x_{1i4t} = 1$, ($t = 1, \dots, 4$), $i = K/2 + 1, \dots, 3K/4$; $x_{1i1t} = x_{1i2t} = x_{1i3t} = x_{1i4t} = 1$, ($t = 1, \dots, 4$), $i = 3K/4 + 1, \dots, K$. The second covariate was chosen as $x_{2ijt} = t/T$, ($t = 1, \dots, 4$), $i = 1, \dots, K$. The within-cluster correlation structure, $C_{1i}(\phi)$, selected was exchangeable (equicorrelation), a natural choice for most practical applications. For the longitudinal correlations we considered AR(1), MA(1) and exchangeable correlation structures.

For selected values of α and ϕ , we computed the relative asymptotic efficiency of $\hat{\beta}_w$ to $\hat{\beta}_Q$ defined by (6.2.9). Results of these computations are reported in Tables 6.1 and 6.2.

In Table 6.1, when the true longitudinal correlation structure is AR(1), the regression parameter β_1 estimated using the longitudinal independence correlation structure is more efficient than the estimator obtained by using the within-cluster independence structure or the total independence correlation structure. On the other hand, the estimator of β_2 , estimated using the within-cluster independence correlation structure is more efficient than the estimator obtained under the longitudinal independence

Table 6.1: *Percentage relative asymptotic efficiency of $\hat{\beta}_W$ to $\hat{\beta}_Q$, the true correlation matrix is $C_i(\phi, \alpha) = C_{1i}(\phi) \otimes C_{2i}(\alpha)$, $C_{1i}(\phi)$ is exchangeable, $C_{2i}(\alpha)$ is AR(1) and $\text{logit}(\mu_{ijt}) = \beta_1 x_{1ijt} + \beta_2 x_{2ijt}$ with $\beta_1 = \beta_2 = 1$.*

True correlation		Total		Within-cluster		Longitudinal	
		Independence		Independence		Independence	
$C_{1i}(\phi) \otimes C_{2i}(\alpha)$		$I_{n_i} \otimes I_T$		$I_{n_i} \otimes C_{2i}(\alpha)$		$C_{1i}(\phi) \otimes I_T$	
$\phi = \text{EXC}$	$\alpha = \text{AR}(1)$	β_1	β_2	β_1	β_2	β_1	β_2
0.20	0.20	88.60	98.18	89.19	99.52	99.23	98.67
	0.50	86.77	90.70	89.15	99.44	97.04	91.22
	0.80	86.37	66.13	89.04	99.14	96.49	66.55
0.50	0.20	67.89	96.99	68.27	98.33	99.24	98.66
	0.50	66.56	89.30	68.17	97.92	97.11	91.05
	0.80	66.34	64.49	67.83	96.77	96.65	65.91

EXC, Exchangeable correlation; AR(1), First-order Autoregression

correlation structure or the total independence correlation structure. It is also clear that for small to moderately large ϕ ($= 0.2, 0.5$), and small to moderately large α ($= 0.2, 0.5$), the longitudinal independence correlation structure is preferable since the efficiency loss in this case ranges from 0 to 9% whereas the efficiency loss by using either the within-cluster correlation structure or the total independence correlation structure ranges from 0 to 35%.

Table 6.2 reports the relative asymptotic efficiency computations for the case when the true longitudinal correlation is MA(1). In this case, we observe that the regression

Table 6.2: *Percentage relative asymptotic efficiency of $\hat{\beta}_W$ to $\hat{\beta}_Q$, the true correlation matrix is $C_i(\phi, \alpha) = C_{1i}(\phi) \otimes C_{2i}(\alpha)$, $C_{1i}(\phi)$ is exchangeable, $C_{2i}(\alpha)$ is MA(1) and $\text{logit}(\mu_{ijt}) = \beta_1 x_{1ijt} + \beta_2 x_{2ijt}$ with $\beta_1 = \beta_2 = 1$.*

True correlation		Total		Within-cluster		Longitudinal	
		Independence		Independence		Independence	
$C_{1i}(\phi) \otimes C_{2i}(\alpha)$		$I_{n_i} \otimes I_T$		$I_{n_i} \otimes C_{2i}(\alpha)$		$C_{1i}(\phi) \otimes I_T$	
$\phi = \text{EXC}$	$\alpha = \text{MA}(1)$	β_1	β_2	β_1	β_2	β_1	β_2
0.20	0.20	88.53	98.32	89.19	99.52	99.17	98.81
	0.40	86.59	94.69	89.18	99.50	96.95	95.18
	0.49	85.31	91.14	89.15	99.48	95.49	92.62
0.40	0.20	67.84	97.17	68.29	98.36	99.18	98.80
	0.40	66.39	93.48	68.28	98.25	96.99	95.13
	0.49	65.43	89.91	68.24	98.15	95.55	91.52

EXC, Exchangeable correlation; MA(1), First-order Moving Average

parameter, β_1 , estimated by using the longitudinal independence correlation structure is more efficient than the estimator resulting by using the within-cluster correlation structure or the total independence correlation structure. On the other hand, the regression parameter, β_2 estimated by using the within-cluster independence structure is more efficient. Overall, however, the use of the longitudinal independence correlation structure in estimating β_1 or β_2 is preferable since the loss of efficiency using this approach ranges from 0 to 9% for ϕ ($= 0.2, 0.4$) and α ($= 0.2, 0.4$). On the other hand, the efficiency loss by using either the within-cluster independence structure or

the total independence is much higher, ranging from 0 to 32%.

We now turn to case (iv) where the working correlation structure is $C_i^*(\phi, \alpha) = C_{1i}(\phi) \otimes C_{2i}^*(\alpha)$. We assume that $\hat{\alpha} \rightarrow \alpha_0 (\neq \alpha)$, where α_0 and ϕ are used in $V_i = A_i^{\frac{1}{2}} C_i^*(\phi, \alpha) A_i^{\frac{1}{2}}$ in equations (6.2.7) and (6.2.8), α and ϕ are used for $\Sigma_i^{-1}(\phi, \alpha)$ in equations (6.2.4) and (6.2.8).

We considered an exchangeable correlation matrix, $C_{1i}(\phi)$ for the true within-cluster correlation, and an AR(1), an MA(1) or an exchangeable correlation matrix, $C_{2i}(\alpha)$ as the true longitudinal correlation, versus different choices of working longitudinal correlation structure $C_{2i}^*(\alpha)$: (1) a true exchangeable within-cluster correlation structure, $C_{1i}(\phi)$ and a true MA(1) longitudinal correlation structure $C_{2i}(\alpha)$, versus a true exchangeable within-cluster correlation structure $C_{1i}(\phi)$ and a working exchangeable longitudinal correlation structure $C_{2i}^*(\alpha)$; (2) a true exchangeable within-cluster correlation structure $C_{1i}(\phi)$ and a true MA(1) longitudinal correlation structure $C_{2i}(\alpha)$, versus a true exchangeable within-cluster correlation structure, $C_{1i}(\phi)$, and a working AR(1) longitudinal correlation structure $C_{2i}^*(\alpha)$; (3) a true exchangeable within-cluster correlation structure $C_{1i}(\phi)$, and a true AR(1) longitudinal correlation structure, $C_{2i}(\alpha)$, versus a true exchangeable within-cluster correlation structure $C_{1i}(\phi)$, and a working exchangeable longitudinal structure, $C_{2i}^*(\alpha)$; (4) a true exchangeable within-cluster correlation structure $C_{1i}(\phi)$, and a true AR(1) longitudinal correlation structure $C_{2i}(\alpha)$, versus a true exchangeable within-cluster correlation structure $C_{1i}(\phi)$, and a working MA(1) longitudinal correlation structure $C_{2i}^*(\alpha)$.

Case (1): Comparing a true MA(1) versus a working exchangeable longitudinal correlations, for a fixed exchangeable within-cluster correlation

For fixed β , a moment estimator, $\hat{\alpha}$, for α under a working exchangeable longitudinal correlation structure is a solution to the estimating equation given by

$$\sum_{i=1}^K \sum_{j=1}^{n_i} \sum_{t \neq t'=1}^T (r_{ijt} r_{ijt'} - \alpha) = 0, \quad (6.2.10)$$

where $r_{ijt} = (y_{ijt} - \mu_{ijt}(\beta)) / \{\text{var}(y_{ijt})\}^{\frac{1}{2}}$. Since the true longitudinal correlation is MA(1), $E(\hat{\alpha}) = 2\alpha/T$. Therefore, if $\hat{\alpha}$ exists, then $\hat{\alpha}$ converges to $\alpha_0 = 2\alpha/T$. To compute the percentage relative asymptotic efficiency of $\hat{\beta}_W$ we use α for $C_{2i}(\alpha)$ in (6.2.4) and (6.2.8), and $\alpha_0 = 2\alpha/T$ for $C_{2i}^*(\alpha)$ in (6.2.7) and (6.2.8). The results of these computations are reported in Table 6.3.

The results in Table 6.3 show that when the true longitudinal correlation structure is MA(1), the estimators $\hat{\beta}_W = (\hat{\beta}_{1W}, \hat{\beta}_{2W})'$, resulting from the use of working longitudinal independence are slightly more efficient (efficiency loss ranges from 0 to 8%) than the estimators obtained by using working exchangeable longitudinal correlation structure (efficiency loss ranges from 0 to 12%).

Case (2): Comparing a true MA(1) versus a working AR(1) longitudinal correlations, for a fixed exchangeable within-cluster correlation

Fix β , a moment estimator, $\hat{\alpha}$, of α under a working AR(1) longitudinal correlation structure, is a solution to the estimating equation

$$\sum_{i=1}^K \sum_{j=1}^{n_i} \sum_{t < t'}^T (r_{ijt} r_{ijt'} - \alpha^{|t-t'|}) = 0. \quad (6.2.11)$$

Since the true longitudinal correlation is MA(1), $E(\sum_{t \leq t'}^T r_{ijt} r_{ijt'}) = (T-1)\alpha$ and so

Table 6.3: *Percentage relative asymptotic efficiency of $\hat{\beta}_W$ to $\hat{\beta}_Q$. Use α in $C_{2i}(\alpha) =$ true MA(1) and α_0 in $C_{2i}^*(\alpha) =$ working EXC or AR(1); $\text{logit}(\mu_{ijt}) = \beta_1 x_{1ijt} + \beta_2 x_{2ijt}$ with $\beta_1 = \beta_2 = 1$.*

True correlation		Longitudinal		Working longitudinal correlation					
EXC	MA(1)	independence		EXC			AR(1)		
ϕ	α	β_1	β_2	α_0	β_1	β_2	α_0	β_1	β_2
0.20	0.20	99.17	98.81	0.10	99.13	98.46	0.18	99.98	99.94
	0.40	96.95	95.18	0.20	96.76	93.03	0.32	99.80	98.87
	0.49	95.49	91.62	0.24	95.20	88.08	0.38	99.54	96.59
0.40	0.20	99.18	98.80	0.10	99.13	98.44	0.18	99.98	99.94
	0.40	96.99	95.13	0.20	96.78	92.96	0.32	99.80	98.85
	0.49	95.55	91.52	0.24	95.22	87.96	0.38	99.54	96.52

EXC, Exchangeable; AR(1), First-order autoregression; MA(1), First-order moving average

from (6.2.11) it is easy to show that α_0 is a solution to

$$(T-1)\alpha - \frac{\alpha_0}{1-\alpha_0} \left\{ T - \frac{1-\alpha_0^T}{1-\alpha_0} \right\} = 0. \quad (6.2.12)$$

Hence, if $\hat{\alpha}$ exists, then $\hat{\alpha}$ converges in probability to α_0 , where α_0 is related to α through (6.2.12). To compute the percentage relative asymptotic efficiencies, we use α for $C_{2i}(\alpha)$ in (6.2.4) and (6.2.8), and α_0 for $C_{2i}^*(\alpha)$ in (6.2.7) and (6.2.8). The results of these calculations are reported in Table 6.3.

The results in Table 6.3 show that when the true longitudinal correlation structure is MA(1), the estimators $\hat{\beta}_W = (\hat{\beta}_{1W}, \hat{\beta}_{2W})'$, resulting from the use of working AR(1)

longitudinal correlation structure are slightly more efficient (efficiency loss ranges from 0 to 3%) than the estimators obtained by using working longitudinal independence correlation structure (efficiency loss ranges from 0 to 8%).

Case (3): Comparing a true AR(1) versus a working exchangeable longitudinal correlations, for a fixed exchangeable within-cluster correlation

For a fixed β , a moment estimator, $\hat{\alpha}$, of α under a working exchangeable longitudinal correlation structure is a solution to the estimating equation given by (6.2.10). Since the true longitudinal correlation is AR(1), $E(\sum_{t \leq t'}^T r_{ijt} r_{ijt'}) = \alpha^{|t-t'|}$ with $-1 < \alpha < 1$. Hence if $\hat{\alpha}$ exists, it converges to $\alpha_0 \geq 1/(T-1)$, where α_0 is given by

$$\alpha_0 = \frac{2\alpha\{T - (1 - \alpha^T)/(1 - \alpha)\}}{T(T-1)(1 - \alpha)}. \quad (6.2.13)$$

To compute the relative asymptotic efficiencies, we use α for $C_{2i}(\alpha)$ in (6.2.4) and (6.2.8), and α_0 for $C_{2i}^*(\alpha)$ in (6.2.7) and (6.2.8). The results of these computations are reported in Table 6.4.

The results in Table 6.4 show that when the true longitudinal correlation structure is AR(1), the estimators $\hat{\beta}_W = (\hat{\beta}_{1W}, \hat{\beta}_{2W})'$, resulting from the use of working exchangeable longitudinal correlation structure are slightly more efficient (efficiency loss ranges from 0 to 6%) than the estimators obtained by using working longitudinal independence correlation structure (efficiency loss ranges from 0 to 8%).

Case (4): Comparing a true AR(1) versus a working MA(1) longitudinal correlations, for a fixed exchangeable within-cluster correlation

Under a working MA(1) longitudinal correlation structure assumption, a moment

Table 6.4: *Percentage relative asymptotic efficiency of $\hat{\beta}_W$ to $\hat{\beta}_Q$. Use α in $C_{2i}(\alpha) = \text{true AR}(1)$ and α_0 in $C_{2i}^*(\alpha) = \text{working EXC or MA}(1)$; $\text{logit}(\mu_{ijt}) = \beta_1 x_{1ijt} + \beta_2 x_{2ijt}$ with $\beta_1 = \beta_2 = 1$.*

True correlation		Longitudinal		Working longitudinal correlation					
EXC	AR(1)	independence		EXC			MA(1)		
ϕ	α	β_1	β_2	α_0	β_1	β_2	α_0	β_1	β_2
0.20	0.20	99.17	98.81	0.11	99.19	98.68	0.20	100.00	99.93
	0.40	96.95	95.18	0.26	97.60	95.51	0.40	99.99	98.64
	0.49	95.49	91.62	0.34	96.93	93.83	0.49	99.90	95.56
0.40	0.20	99.18	98.80	0.11	99.19	98.68	0.20	100.00	99.93
	0.40	96.99	95.13	0.26	97.62	95.48	0.40	99.97	98.61
	0.49	95.55	91.52	0.34	96.96	93.79	0.49	99.91	95.46

EXC, Exchangeable; AR(1), First-order autoregression; MA(1), First-order moving average

estimator, $\hat{\alpha}$, of α is a solution to the estimating equation

$$\sum_{i=1}^K \sum_{j=1}^{n_i} \sum_{t=1}^{T-1} (r_{ijt} r_{ij(t+1)} - \alpha) = 0. \quad (6.2.14)$$

If $\hat{\alpha}$ exists, then in this case $\hat{\alpha}$ converges to $\alpha_0 = \alpha$, because, under the true AR(1) correlation structure, $E(r_{ijt} r_{ij(t+1)}) = \alpha$. Note that, although in the present case α can take any value from -1 to 1, we can use only the range $-0.5 < \alpha (= \alpha_0) < 0.5$ for the efficiency computation. Results of these calculations are reported in Table 6.4.

The results in Table 6.4 show that when the true longitudinal correlation structure is AR(1), the estimators $\hat{\beta}_W = (\hat{\beta}_{1W}, \hat{\beta}_{2W})'$, resulting by using working MA(1) lon-

gitudinal correlation structure are slightly more efficient (efficiency loss ranges from 0 to 4%) than the estimators obtained by using working longitudinal independence correlation structure (efficiency loss ranges from 0 to 8%).

To summarize the results from the efficiency calculations, we note that, efficiency loss in using a working longitudinal independence correlation structure, may be quite minimal for most practical purposes. On the other hand, the use of a working within-cluster independence correlation structure in estimating regression parameters, can lead to substantial loss in efficiency.

6.3 Total Independence Estimating Equations and Informative Cluster Sizes

For the assume model (6.2.1), and under a working covariance matrix V_i , a consistent estimator, $\hat{\beta}_W$, of β is obtained by solving the estimating equations (6.2.5). Under a total working independence assumption, (6.2.5) reduces to

$$U_I(\beta) = \sum_{i=1}^K \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}} \frac{x_{ijt}(y_{ijt} - \mu_{ijt}(\beta))}{a_{ijt}} \frac{\partial \mu_{ijt}}{\partial \eta_{ijt}} = \sum_{i=1}^K \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}} u_{ijt}(\beta) = 0, \quad (6.3.15)$$

where $a_{ijt} = \text{var}(y_{ijt})$. Equation (6.3.15) is based on the working assumption that repeated observations for units in a cluster are independent and cluster units are also independent. There is also an implicit assumption that cluster sizes are noninformative.

As noted in Section 6.2, the solution to (6.3.15) is obtained by iteration and at convergence leads to the estimator, $\hat{\beta}_I$, and the observed information matrix, $J_I(\hat{\beta}_I)$,

which is the value of

$$J(\beta) = -\frac{\partial U_I(\beta)}{\partial \beta'} = \sum_{i=1}^K \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}} \frac{\mathbf{x}_{ijt} \mathbf{x}_{ijt}'}{a_{ijt}} \left(\frac{\partial \mu_{ijt}}{\partial \eta_{ijt}} \right)^2 \quad (6.3.16)$$

evaluated at $\beta = \hat{\beta}_I$.

Under cluster size ignorability, $\hat{\beta}_I$ is consistent for β as $K \rightarrow \infty$. Furthermore, a consistent estimator of the covariance matrix of $\hat{\beta}_I$ is

$$\widehat{V}_I = \{J_I(\hat{\beta}_I)\}^{-1} \left[\sum_{i=1}^K \left\{ \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}} \mathbf{u}_{ijt}(\hat{\beta}_I) \right\} \left\{ \sum_{i=1}^K \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}} \mathbf{u}_{ijt}(\hat{\beta}_I) \right\}' \right] \{J_I(\hat{\beta}_I)\}^{-1} \quad (6.3.17)$$

The estimator $\hat{\beta}_I$ may not be consistent if the cluster sizes are nonignorable, as shown in the example below.

Example 6.1 Nested error linear regression model

Consider a nested error linear regression model

$$y_{ijt} = \mathbf{x}_{it}'\beta + \alpha_i + \epsilon_{ijt}, \quad t = 1, \dots, T_i; \quad j = 1, \dots, n_i; \quad i = 1, \dots, K, \quad (6.3.18)$$

where β is a $p \times 1$ vector of regression parameters, $\mathbf{x}_{it}$ is a $p \times 1$ vector of fixed covariates of the t -th repeated observation in the i -th cluster, y_{ijt} is the response for the t -th repeated observation of the j -th unit in the i -th cluster, α_i is the i -th cluster effect and $n_i = h(\alpha_i)$ is the i -th cluster size. We further assume that the random effects, α_i and the errors ϵ_{ijt} are independent and identically distributed with mean 0 and variances σ_α^2 and σ_ϵ^2 respectively. The model (6.3.18) leads to the marginal model $E(y_{ijt}) = \mu_{it} = \mathbf{x}_{it}'\beta$.

The total independence estimating equations estimator of β for the model $E(y_{ijt}) = \mathbf{x}_{it}'\beta$ is given by

$$\hat{\beta}_I = \left(\sum_{i=1}^K n_i \sum_{t=1}^{T_i} \mathbf{x}_{it} \mathbf{x}_{it}' \right)^{-1} \sum_{i=1}^K \sum_{j=1}^{n_i} \sum_{t=1}^{T_i} \mathbf{x}_{it} y_{ijt}. \quad (6.3.19)$$

As $K \rightarrow \infty$, $E(\hat{\beta}_I)$ converges to

$$\bar{\beta} = \beta + \frac{E\{\alpha h(\alpha)\}}{E\{h(\alpha)\}} \left(\sum_{i=1}^K \sum_{t=1}^{T_i} \mathbf{x}_{it} \mathbf{x}_{it}' \right)^{-1} \sum_{i=1}^K \sum_{t=1}^{T_i} \mathbf{x}_{it}. \quad (6.3.20)$$

Since in general the last term of (6.3.20) is not a zero vector, $\hat{\beta}_I$ is asymptotically biased for β . However, for models with an intercept term β_0 , we have

$$\left(\sum_{i=1}^K \sum_{t=1}^{T_i} \mathbf{x}_{it} \mathbf{x}_{it}' \right)^{-1} \sum_{i=1}^K \sum_{t=1}^{T_i} \mathbf{x}_{it} = (1, 0, \dots, 0)'. \quad (6.3.21)$$

This means that when the assumption of cluster size ignorability fails, inferences based on the total independence estimating equations may be asymptotically invalid.

6.4 Inverse-Sampling Estimating Equations

As discussed in the previous section, methods that implicitly assume that cluster sizes are ignorable (e.g., a total independence estimating equations, TIEE) can lead to asymptotically invalid inferences when this assumption fails. We propose three methods that are asymptotically valid whether cluster sizes are ignorable or informative. These methods are (i) separate longitudinal independence estimating equations (SLIEE); (ii) combined longitudinal independence estimating equation (CLIEE), and (iii) mean longitudinal independence estimating equation (MLIEE). We note that SLIEE and CLIEE methods are repeated subsampling (inverse-sampling) methods, and MLIEE is not a subsampling method, it is motivated by the CLIEE method.

In Subsection 6.4.1, we introduce the separate longitudinal independence estimating equations method. The properties of the combined longitudinal independence

estimating equation method are studied in Subsection 6.4.2, and in Subsection 6.4.3 we introduce the mean longitudinal independence estimating equation method.

6.4.1 Separate Longitudinal Independence Estimating Equations

For this subsampling method, we simply select one unit at random from each cluster to obtain K independent units, where the i -th selected unit has T_{ij} , $j = 1, \dots, n_i$ repeated observations. This process is repeated g times. The resulting subsamples $s_j^* = \{(x_{ijt}, y_{ijt}); t = 1, \dots, T_{ij}; i = 1, \dots, K; j = 1, \dots, g\}$ are conditionally independent (across units) given the original sample s . The inverse-sampling estimators, $\hat{\beta}_j^*$, are obtained by solving g independence estimating equations separately:

$$\begin{aligned} U_j^*(\beta) &= \sum_{i=1}^K \sum_{t=1}^{T_{ij}} \left(\frac{\partial \mu_{ijt}}{\partial \beta} \right)' \frac{(y_{ijt}^* - \mu_{ijt}^*)}{a_{ijt}^*} = \sum_{i=1}^K \sum_{t=1}^{T_{ij}} u_{ijt}^*(\beta) \\ &= \sum_{i=1}^K u_{ij}^*(\beta) = 0, \quad j = 1, \dots, g \end{aligned} \quad (6.4.22)$$

where $\mu_{ijt}^* = E(y_{ijt}^*)$ and $a_{ijt}^* = \text{var}(y_{ijt}^*)$.

In general, (6.4.22) is solved iteratively to obtain the solution $\hat{\beta}_j^*$. The Newton-Raphson's method is commonly used to solve (6.4.22). We note that we require g iterative solutions to get the estimators, $\hat{\beta}_1^*, \dots, \hat{\beta}_g^*$. Hence it could be computationally intensive. The separate longitudinal independence estimating equations estimator of β is given by

$$\hat{\beta}_g = \frac{1}{g} \sum_{j=1}^g \hat{\beta}_j^*. \quad (6.4.23)$$

The inverse-sampling covariance matrix of $\hat{\beta}_g$ is given by

$$\widehat{V}_g = \frac{1}{g} \sum_{j=1}^g \widehat{V}_j^* - \frac{1}{g} \sum_{j=1}^g (\hat{\beta}_j^* - \hat{\beta}_g)(\hat{\beta}_j^* - \hat{\beta}_g)', \quad (6.4.24)$$

where $\widehat{V}_j^*$ is the linearization variance estimator of $\hat{\beta}_j^*$:

$$\widehat{V}_j^* = \{J_j^*(\hat{\beta}_j^*)\}^{-1} \left\{ \sum_{i=1}^K u_{ij}(\hat{\beta}_j^*) u_{ij}(\hat{\beta}_j^*)' \right\} \{J_j^*(\hat{\beta}_j^*)\}^{-1}, \quad (6.4.25)$$

where $J_j^*(\hat{\beta}_j^*)$ is $J_j^*(\beta) = -\partial U_j^*(\beta) / \partial \beta'$ evaluated at $\hat{\beta}_j^*$.

For small and moderately large number of clusters, the variance estimator of $\hat{\beta}_g$ could be unstable, by producing negative values. An alternative approach that may be less computationally intensive and less likely to produce negative variance estimates, when the number of clusters is small, is the combined longitudinal independence estimating equations. The properties of this method are studied in the next subsection.

6.4.2 Combined Longitudinal Independence Estimating Equation

The method of subsampling in this method is the same as in Subsection 6.4.1, namely, select one unit at random from each cluster, where the i -th ($i = 1, \dots, K$) selected unit has T_{ij} , $j = 1, \dots, n_i$ repeated observations. The combined longitudinal independence estimating equation (CLIEE) is obtained by combining the g equations in (6.4.22) to get

$$U_{gc}(\beta) = \frac{1}{g} \sum_{j=1}^g U_j^*(\beta) = \frac{1}{g} \sum_{j=1}^g \sum_{i=1}^K u_{ij}^*(\beta) = 0. \quad (6.4.26)$$

In general, we solve (6.4.26) using the Newton-Raphson iterations (6.2.3) with $U(\hat{\beta}^{(r-1)})$ changed to $U_{gc}(\beta^{(r-1)})$ and $J(\hat{\beta}^{(r-1)})$ changed to $J_{gc}(\hat{\beta}^{(r-1)})$, where

$$J_{gc}(\beta) = -\frac{\partial U_{gc}(\beta)}{\partial \beta^T} = \frac{1}{g} \sum_{j=1}^g \sum_{i=1}^K \sum_{t=1}^{T_{ij}} \frac{1}{a_{ijt}^*} \left(\frac{\partial \mu_{ijt}}{\partial \beta} \right)' \left(\frac{\partial \mu_{ijt}}{\partial \beta} \right). \quad (6.4.27)$$

At convergence, we get the estimator $\hat{\beta}_{gc}$ as well as the observed information matrix $J_{gc}(\hat{\beta}_{gc})$. We note that we solve the combined longitudinal independence estimating equation only once to get $\hat{\beta}_{gc}$, whereas we solve g separate within-cluster independence estimating equations to obtain $\hat{\beta}_1^*, \dots, \hat{\beta}_g^*$ and then averaging the estimators to obtain the separate independence estimating equations estimator $\hat{\beta}_g$ given in Subsection 6.4.1.

Theorems 6.1 and 6.2 below summarize some properties of the combined longitudinal independence estimating equation function.

Theorem 6.1 1. *Conditional on the original sample s , $U_{gc}(\beta)$ converges to*

$$E\{U_{gc}(\beta)|s\} = U_M(\beta), \text{ as } g \rightarrow \infty, \text{ where}$$

$$U_M(\beta) = \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}} \left(\frac{\partial \mu_{ijt}}{\partial \beta} \right)' \frac{(y_{ijt} - \mu_{ijt})}{a_{ijt}} = \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} u_{ij}(\beta).$$

$$2. E\{U_{gc}(\beta)\} = E\{U_1^*(\beta)\}.$$

3. *Let $\hat{\beta}_M$ be a solution to $U_M(\beta) = 0$ and $\hat{\beta}_{gc}$ a solution to $U_{gc}(\beta) = 0$. Then under mild regularity conditions, $\hat{\beta}_{gc}$ and $\hat{\beta}_M$ are asymptotically equivalent.*

Proof

1. Conditional on s , $U_1^*, \dots, U_g^*$ are independent identically distributed bounded random vectors, hence the result follows.

2. By the standard relationship between conditional and unconditional expectations we have

$$E\{U_{gc}(\beta)\} = E[E\{U_1^*(\beta)|s\}] = E\{U_1^*(\beta)\}.$$

3. Consider the following regularity conditions:

C1. For some $\delta > 0$,

$$\sup_i E \left\| \frac{\partial \mathbf{u}_{ij}(\beta)}{\partial \beta'} \right\|^{1+\delta} < \infty.$$

C2. Assume that

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{i=1}^K E \left\{ -\frac{\partial}{\partial \beta'} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\beta) \right\} = \mathbf{I}(\beta),$$

where $\mathbf{I}(\beta)$ exists and it is nonsingular.

C3. For some $\delta > 0$,

$$\sup_i E \|\mathbf{u}_{ij}(\beta)\|^{2+\delta} < \infty.$$

C4. Consider

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{i=1}^K \text{Var} \{\mathbf{u}_i(\beta)\} = \mathbf{V}_U,$$

a positive-definite covariance matrix and $\mathbf{u}_i(\beta) = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\beta)$.

Under assumptions C1, C2 and using the Taylor expansion of $U_M(\hat{\beta}_M)$ around β we have,

$$\hat{\beta}_M - \beta = \{\mathbf{I}(\beta)\}^{-1} \frac{1}{K} U_M(\beta) + o_p(1/\sqrt{K}). \quad (6.4.28)$$

Similarly,

$$\hat{\beta}_{gc} - \beta = \{I(\beta)\}^{-1} \frac{1}{K} U_{gc}(\beta) + o_p(1/\sqrt{K}). \quad (6.4.29)$$

From (6.4.28) and (6.4.29) we have

$$\|\hat{\beta}_M - \hat{\beta}_{gc}\| \leq \left\| \{I(\beta)^{-1}\} \right\| \frac{1}{K} \|U_M(\beta) - U_{gc}(\beta)\| + o_p(1/\sqrt{K}). \quad (6.4.30)$$

By Result 1 of Theorem 6.1, the first term of the right hand side of (6.4.30) converges to zero in probability as $g \rightarrow \infty$ and the result follows as $K \rightarrow \infty$. ■

Theorem 6.2 *The variance of $U_{gc}(\beta)$ may be expressed in the following two ways*

1. $\text{var}\{U_{gc}(\beta)\} = \text{var}\{U_M(\beta)\} + \frac{1}{g} E[\text{var}\{U_1^*(\beta)|s\}].$
2. $\text{var}\{U_{gc}(\beta)\} = \text{var}\{U_1^*(\beta)\} - \frac{g-1}{g} E[\text{var}\{U_1^*(\beta)|s\}].$

Proof

1. Using the conditional and unconditional variances and noting that given s the U_j^* 's are conditionally independent we have

$$\begin{aligned} \text{var}\{U_{gc}(\beta)\} &= \text{var}[E\{U_{gc}(\beta)|s\}] + E[\text{var}\{U_{gc}(\beta)|s\}] \\ &= \text{var}\{U_M(\beta)\} + \frac{1}{g} E[\text{var}\{U_1^*(\beta)|s\}]. \end{aligned}$$

2. Apply Result 3 of Theorem 6.1 with $g = 1$ to obtain

$$\text{var}\{U_M(\beta)\} = \text{var}\{U_1^*(\beta)\} - E[\text{var}\{U_1^*(\beta)|s\}],$$

and substituting into Result 3 of Theorem 6.1, the result follows. ■

6.4.3 Mean Longitudinal Independence Estimating Equation

The mean longitudinal independence estimating equation is given by

$$\begin{aligned} U_M(\beta) &= \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{t=1}^{T_{ij}} \left(\frac{\partial \mu_{ijt}}{\partial \beta} \right)' \frac{(y_{ijt} - \mu_{ijt})}{a_{ijt}} \\ &= \sum_{i=1}^K \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\beta) = 0. \end{aligned} \quad (6.4.31)$$

It is worth noting that the mean longitudinal independence estimating equation method is not an inverse-sampling method. For this reason, it is not as computationally intensive as the inverse-sampling methods discussed in Subsections 6.4.1 and 6.4.2.

In general, we solve (6.4.31) using Newton-Raphson iterations (6.2.3) with $U(\hat{\beta}^{(r-1)})$ changed to $U_M(\hat{\beta}^{(r-1)})$ and $J(\hat{\beta}^{(r-1)})$ changed to $J_M(\hat{\beta}^{(r-1)})$, where $J_M(\beta) = -\partial U_M / \partial \beta'$. At convergence, we get the estimator $\hat{\beta}_M$ as well as the observed information matrix $J_M(\hat{\beta}_M)$.

Theorem 6.3 below states an asymptotic normality property for $\hat{\beta}_M$

Theorem 6.3 *Under mild regularity conditions, and given that $\hat{\beta}_M$ is a consistent estimator of β , then $K^{1/2}(\hat{\beta}_M - \beta)$ is asymptotically multivariate normal with zero mean and covariance matrix consistently estimated by*

$$\widehat{V}_M = K \{J_M(\hat{\beta}_M)\}^{-1} \widehat{V}_{U_M} \{J_M(\hat{\beta}_M)\}^{-1}, \quad (6.4.32)$$

where $J_M(\hat{\beta}_M)$ is $-\partial U_M(\beta) / \partial \beta'$ evaluated at $\beta = \hat{\beta}_M$ and

$$\widehat{V}_{U_M} = \widehat{V} \{U_M(\hat{\beta}_M)\} = \sum_{i=1}^K \left[\left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{u}_{ij}(\hat{\beta}_M) \right\} \left\{ \frac{1}{n_i} \sum_{l=1}^{n_i} \mathbf{u}_{il}(\hat{\beta}_M) \right\}' \right].$$

Proof

By assumptions C3 and C4, and for any vector $\mathbf{a}$,

$$\begin{aligned} \frac{\sum_{i=1}^K E |\mathbf{a}' \mathbf{u}_i(\boldsymbol{\beta})|^{2+\delta}}{\left\{ \sum_{i=1}^K \text{Var}(\mathbf{a}' \mathbf{u}_i(\boldsymbol{\beta})) \right\}^{\frac{2+\delta}{2}}} &= O(1/K^{\delta/2}) \\ &\rightarrow 0 \text{ as } K \rightarrow \infty. \end{aligned}$$

Hence by Lyapunov's central limit theorem

$$\frac{1}{\sqrt{K}} \mathbf{U}_M(\boldsymbol{\beta}) = \frac{1}{\sqrt{K}} \sum_{i=1}^K \mathbf{u}_i(\boldsymbol{\beta}) \rightarrow_d N(\mathbf{0}, \mathbf{V}_U). \quad (6.4.33)$$

Using Taylors expansion of $\mathbf{U}_M(\hat{\boldsymbol{\beta}}_M)$ around $\boldsymbol{\beta}$, assumption C2, and (6.4.33), we have

$$\begin{aligned} \sqrt{K}(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}) &= \left\{ -\frac{1}{K} \sum_{i=1}^K \frac{\partial \mathbf{u}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} \right\}^{-1} \left\{ \frac{1}{\sqrt{K}} \sum_{i=1}^K \mathbf{u}_i(\boldsymbol{\beta}) \right\} \\ &\rightarrow N(\mathbf{0}, \mathbf{I}(\boldsymbol{\beta})^{-1} \mathbf{V}_U \mathbf{I}(\boldsymbol{\beta})^{-1}), \end{aligned}$$

where $\tilde{\boldsymbol{\beta}} \in (\hat{\boldsymbol{\beta}}_M, \boldsymbol{\beta})$. Now, note that $\mathbf{I}(\boldsymbol{\beta})$ is consistently estimated by $\hat{\mathbf{J}}(\hat{\boldsymbol{\beta}}) = \mathbf{J}_M(\hat{\boldsymbol{\beta}})/K = -\partial \mathbf{U}_M(\boldsymbol{\beta})/K \partial \boldsymbol{\beta}'$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\mathbf{V}_U$ is consistently estimated by $\widehat{\mathbf{V}}_{U_M}/K$, to get (6.4.32). ■

The SLIEE, CLIEE and MLIEE can be modified so that a working longitudinal correlation is used as oppose to a working longitudinal independence as discussed extensively in this chapter.

Chapter 7

Conclusions and Future Research

7.1 Conclusions

In Part I of the thesis we studied analysis of complex survey data using inverse sampling methods. We studied estimation of simple population parameters (e.g., a population total) under both exact and approximate inverse sampling methods. We showed that under mild regularity conditions (including existence of an exact inverse sampling scheme), the inverse-sampling estimator, $\hat{\theta}_{gc}$ and the associated variance estimator $\hat{V}(\hat{\theta}_{gc})$ converge to the corresponding full-sample estimator, $\hat{\theta}$ and variance estimator, $\hat{V}(\hat{\theta})$ respectively, as $g \rightarrow \infty$.

For complex census population parameters, such as linear regression and logistic regression parameters, we proposed a combined estimating equation (CEE) and demonstrated its advantages over separate estimating equations (SEE). We also showed that under mild regularity conditions, the combined estimating equation estimator, $\hat{\theta}_{gc}$ converges to the full-sample estimator, $\hat{\theta}$ as $g \rightarrow \infty$; the associated inverse-

sampling variance estimator $\hat{V}_L(\hat{\theta}_{gc})$, converges to the full-sample variance estimator, $\hat{V}_L(\hat{\theta})$. We also demonstrated that, the CEE estimator leads to asymptotically valid inferences.

Applications of the CEE method were demonstrated in poststratification estimation of a population total and census population nonlinear parameters. We studied inverse-sampling Wald and Pearson tests for goodness-of-fit problems in complex survey context. For inferences on model parameters, we studied properties of an inverse-sampling quasi-score test.

Part II of the thesis examined a number of methods to analyze correlated biological data when cluster sizes are ignorable or informative.

We demonstrated in Chapter 4 that methods that implicitly assume that cluster sizes are ignorable, can lead to asymptotically invalid inferences when this assumption fails. We studied alternative methods, combined estimating equation and mean estimating equation that are asymptotically valid when cluster sizes are ignorable or informative. We evaluated the performance of the proposed methods both analytically and through simulation studies. Properties of a test for cluster sizes ignorability were also studied.

In Chapter 5, we studied properties of the mean conditional estimating equation method as an alternative method to analyze cluster-correlated retrospective data. We showed that this method leads to asymptotically valid inferences whether the correlation structures arise from the random effects or by other unmeasured factors (where the CLR may fail).

Chapter 6 extended the results of Chapter 4 to study analysis of data with two

types of correlation. We studied the efficiency of working correlation-based estimation methods when cluster sizes are ignorable. When cluster sizes are informative, we demonstrated that estimators of the regression parameters based on methods that implicitly assume that cluster sizes are ignorable (e.g., total independence estimating equations, TIEE) can lead to appreciable bias. We studied a number of alternative methods, separate longitudinal independence estimating equations (SLIEE), combined longitudinal independence estimating equations (CLIEE) and mean longitudinal independence estimating equations (MLIEE) that lead to asymptotically valid inferences when cluster sizes are ignorable or informative.

7.2 Future Research

Parts I and II of this thesis studied a number of inverse-sampling methods to analyze complex survey data and cluster-correlated biological data respectively.

Part I demonstrated the potential of inverse sampling in providing public-use micro data files, consisting of multiple simple random subsamples, that can be used to make valid inferences, such as regression and categorical data analysis. However, further work is needed to provide necessary enhancements to standard software in order to implement the CEE method in practice for standard error calculations.

We studied inverse sampling algorithms for a number of standard complex sample designs. However, work still remains to be done to cover other complex sampling designs used in more complex surveys such as in the Labour Force Survey by Statistics Canada.

In the context of point estimation of a total, Y , full-sample weights are sometimes adjusted to avoid problems with variance inflation due to influential observations. An interesting area of investigation is how to modify the inverse sampling algorithms so that the inverse-sampling estimator, $\tilde{Y}_g$ converges to the adjusted-weight full-sample estimator say, $\tilde{Y}$, as $g \rightarrow \infty$.

There has been a number of panel surveys such as the National Health and Nutrition Examination Survey Epidemiology Follow-up Study (Folson, La Vange and Williams, 1989) where design-based proportional hazards model is a suitable tool of analysis. Binder (1992) proposed fitting Cox's proportional hazards models to this type of survey data. Lin (2000) established the asymptotic properties of Binder's method. This is another area where we propose to investigate the use of inverse sampling CEE method.

Part II of this thesis focused primarily on the analysis of cluster-correlated biological data where cluster sizes may be ignorable or nonignorable. In Chapter 5, we discussed analysis of clustered biological response data, in particular, data from clustered matched case-control studies. We showed that when the correlation structure is generated by the random effects of a mixed-effects logistic regression model, then both the CLR and the MCEE methods lead to asymptotically valid inferences. However, as was conjectured, CLR can lead to invalid inferences when this assumption fails. More work, in particular through simulation studies, is needed to further assess the performance of the proposed method and the CLR approach.

Appendix I - Splus Syntax to Implement Mean Estimating Equation

```
sandwich ← function(formula, family=fam, data)
{
  # data: data frame of input data must have
  # a variable clusterID, giving cluster labels to
  # any S-type model formula
  cluster.sizes ← table(factor(data$clusterID))
  result ← glm(formula, family=fam, weights=
1/cluster.sizes[factor(data$clusterID)], data=data)
  X ← model.matrix(result)
  W ← results$weights
  I.beta ← t(X)%*% (W*X)
  U ← X*residuals(result, type="response")
  Ui ← apply(U,2 function(X)tapply(X,clusterID, mean))
  IB.inv ← solve(I.beta)
  temp.mat ← IB.inv %*% t(Ui)
  #
  # Output: usual glm object, plus sandwich variance estimator
  list(glm.obj = result, sandwich=temp.mat%*% t(temp.mat))
}
```

Appendix II - Splus Syntax to Implement CLR for Cluster-correlated Matched Case-control Data

```
# data:  $(y, x_1, \dots, x_p)$ 

# y - response variable taking the value 1 (case) or 0 (control)

#  $(x_1, \dots, x_p)$  a  $p \times 1$  vector of covariates

# clusterid - cluster identification

#

data1 ← list(time=rep(1,length(y)),  $x_1, \dots, x_p, y$ , clusterid)

output ← coxph(Surv(time,y) ~  $x_1 + x_2 + \dots + x_p$  + strata(clusterid),

method='exact', data=data1)

# coxph(...) solves exactly the same log-likelihood function as the matched

# logistic, and thus gives the correct estimators of the regression parameters,

# standard errors, residuals, etc.
```


References

- Altham, P. A. E. (1976). Discrete variable analysis for individuals grouped into families, *Biometrika*, **63**, 263-269.
- Battese, G. E., Harter, M. R. and Fuller, A. W. (1988). An error-component model for prediction county crop areas using survey and satellite data, *Journal of the American Statistical Association*, **83**, 28-36.
- Best, J. A., Brown, K. S., Cameron, R., Manske, S. M. and Santi, S. (1995). Gender and predisposing attributes as predictors of smoking onset: Implications for theory and practice, *Journal of Health Education*, **26**, S52-S60.
- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys, *International Statistical Review*, **51**, 279-292.
- Binder, D. A. (1992). Fitting Cox's proportional hazard models from survey data, *Biometrika*, **79**, 139-147.
- Brier, S. S. (1978). Categorical data models for complex data structures, Unpublished Ph.D. thesis, University of Minnesota, School of Statistics.
- Brown, K. S. and Cameron, R. (1997). Long-term evaluation of an elementary and secondary school smoking intervention. Final Report to the National Health Research and Development Program (Canada).
- Cochran, W. G. (1977). *Sampling Techniques*. 3rd ed. New York: Wiley.

- Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Crowder, M. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measures, *Biometrika*, **82**, 407-410.
- Fellegi, I. P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples, *Journal of the American Statistical Association*, **71**, 665-674.
- Gansky, S. A., Weintraub, J. A. and Shain, S. (1999). MultiPied investigators: Family aggregation of periodontal status in a two generation cohort, *Journal of Dental Research*, **78**, 123-132.
- Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation, *International Statistical Review*, **54**, 127-138.
- Hidiroglou, M. A., Rao, J. N. K. and Yung, W. (1999). Variance computation for complex surveys using estimating equations, *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 1999, 3-9.
- Hinkins, S., Oh, H. L. and Scheuren, F. (1997). Inverse sampling design algorithms, *Survey Methodology*, **23**, 11-21.
- Hoffman, E. B. (1998). Within-cluster resampling, unpublished Ph.D. thesis, University of North Carolina, Chapel Hill, United States of America.

- Hoffman, E. B., Sen, P. K. and Weinberg, C. R. (2001). Within-cluster resampling, *Biometrika*, **88**, 1121-1134.
- Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.
- Koch, G. G., Freeman, D. H. and Freeman, J. L. (1975). Strategy in the multivariate analysis of data from complex surveys, *International Statistical Review*, **43**, 59-78.
- Kovacevic, M. S. and Binder, D. A. (1997). Variance estimation for measures of income inequality and polarization - the estimating equations approach, *Journal of Official Statistics*, **13**, 41-58.
- Krewski, D. R. and Rao, J. N. K. (1981). Inference from stratified samples: Properties of Linearization, Jackknife and Balance Repeated Replication Methods, *The Annals of Statistics*, **9**, 1010-1019.
- Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13-22.
- Nathan, G. (1975). Tests of independence in contingency tables from stratified proportional samples. *Sankhyā C*, **37**, 77-87.
- Rao, J. N. K. and Scott, A. J. (1979). Chi-squared tests for analysis of categorical data from complex surveys, *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 1979, 58-66.

- Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness of fit and independence on two-way tables, *Journal of the American Statistical Association*, **76**, 221-230.
- Rao, J. N. K. and Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data, *Annals of Statistics*, **12**, 46-60.
- Rao, J. N. K. and Scott, A. J. (1992). A simple method for the analysis of clustered binary data, *Biometrics*, **48**, 577-585.
- Rao, J. N. K. and Scott, A. J. (1999). A simple method for analysing overdispersion in clustered Poisson data, *Statistics in Medicine*, **18**, 1373-1385.
- Rao, J. N. K., Scott, A. J. and Skinner, C. J. (1998). Quasi-score tests with survey data, *Statistica Sinica*, **8**, 1059-1070.
- Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika*, **79**, 811-822.
- Rao, J. N. K. and Shao, J. (1999). Modified balanced repeated replication for complex survey data, *Biometrika*, **86**, 403-415.
- Rieger, R. H. and Weinberg, C. R. (2002). Analysis of clustered binary outcomes using within-cluster paired resampling, *Biometrics*, **58**, 332-341.
- Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data, *Journal of the American Statistical Association*, **83**, 231-241.

- Särndal, C. E., Swensson, B. and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag: New York.
- Shao, J. (1999). *Mathematical Statistics*. Springer-Verlag: New York.
- Shuster, J. J. and Downing, D. J. (1976). Two-way contingency tables from complex sampling schemes, *Biometrika*, **63**, 271-276.
- Skinner, C. J., Holt, D. and Smith, T. M. F. (Eds.) (1989), *Analysis of Complex Surveys*. New York: Wiley.
- Sutradhar, B. C. and Das, K. (1999). On the efficiency of regression estimators in generalised linear models for longitudinal data, *Biometrika*, **86**, 459-465.
- Thomas, D. R. and Rao, J. N. K. (1987). Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling, *Journal of the American Statistical Association*, **82**, 630-636.
- Williamson, J. M., Datta, S. and Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative, *Biometrika*, **59**, 36-42.
- Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag: New York.