CAUSAL INFERENCE WITH COMPLEX SAMPLING DESIGNS

By

ZHULIN HE

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2012

© 2012 Zhulin He

To my family, who offered me unconditional love and support throughout the course of my Ph.D. study

ACKNOWLEDGMENTS

I would like to express my sincere thanks to many individuals who offered their invaluable assistance and generous help in the preparation of this dissertation.

I am heartily thankful to my advisor, Dr. Babette Brumback, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of biostatistcs, especially in causal modeling and inference and complex survey data analysis. During my four year Ph.D. study, she not only mentored me how to do research in statistics/biostatistics, but also gave me generous help and suggestions on my further career. I consider myself very lucky to have been studying under her guidance, and I will always be thankful for everything I learned from her.

I want to thank the Department of Biostatistics for giving me all the chances and facilities to my graduate study. The faculties and staffs are very nice to the graduate students and always give me a lot of advices and help for my graduate study and life in Gainesville. I really enjoy the peaceful and great research environment in the Department of Biostatistics at the University of Florida.

I would like to thank the members of my doctoral committee for their input, valuable discussions and accessibility. In particular, I would like to thank Dr. Xiaomin Lu and Dr. Amy Cantrell, who gave me advices and help in my graduate study, encouraged me for my further career and recommended me without reservation. I would also like to thank Dr. Richard Rheingans, whose enthusiastic instructions and generous encouragements have facilitated me much through my research work with Dr. Babette Brumback in SWASH project.

Finally, and most importantly, I would express my thanks to my parents and my husband who always give me their endless love and boundless support. Thank you for giving me great confidence and strong support in my life.

Chapter 1 is reprinted from Statistics in Medicine, Volume 29, Issue 18, Brumback BA, Dailey AB, He Z, Brumback LC, and Livingston MD, “Efforts to Adjust for Confounding by Neighborhood Using Complex Survey Data”, pages 1890-1899, 2010, with permission from John Wiley & Sons.

Chapter 2 is partially reprinted from Communications in Statistics – Theory and Methods, He Z, Brumback BA, “An Equivalence of Conditional and Unconditional Maximum Likelihood Estimators via Infinite Replication of Observations”, in press, with permission from Taylor & Francis Group.

Chapter 3 is reprinted from Statistics in Medicine, Volume 30, Issue 9, Brumback BA, He Z, “Adjusting for confounding by neighborhood using complex survey data”, pages 965-972, 2011, with permission from John Wiley & Sons.

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS 4

LIST OF TABLES 8

LIST OF FIGURES 9

ABSTRACT 10

CHAPTER

1 EFFORTS TO ADJUST FOR CONFOUNDING BY NEIGHBORHOOD USING COMPLEX SURVEY DATA 12

1.1 Outline 12
1.2 Introduction 12
1.3 Motivating NHIS Example 15
1.4 Modeling Framework 16
1.5 Approaches to Estimation with Complex Survey Data 18
1.5.1 Ordinary Logistic Regression for Complex Survey Data 18
1.5.2 Conditional Logistic Regression for Complex Survey Data 19
1.5.3 GLMM Regression for Complex Survey Data 21
1.6 Simulation Study 24
1.6.1 Ordinary Logistic Regression for Complex Survey Data 25
1.6.2 Conditional Logistic Regression for Complex Survey Data 25
1.6.3 GLMM Regression for Complex Survey Data 26
1.6.4 Results 26
1.7 Application to NHIS Data 30
1.8 Discussion 32

2 AN EQUIVALENCE OF CONDITIONAL AND UNCONDITIONAL MAXIMUM LIKELIHOOD ESTIMATORS VIA INFINITE REPLICATION OF OBSERVATIONS 33

2.1 Outline 33
2.2 Introduction 33
2.3 Main Results 37
2.4 Simulation Study 46
2.5 Discussion 48

3 ADJUSTING FOR CONFOUNDING BY NEIGHBORHOOD USING COMPLEX SURVEY DATA 50

3.1 Outline 50
3.2 Introduction 50
3.3 Motivating NHIS Example 52
3.4 Modeling Framework 53

3.5 Estimation with Complex Survey Data . . . 54
3.5.1 Simplified Implementation . . . 56
3.5.2 Extension to Multiple Individual-Level Covariates . . . 58
3.5.3 Specifying the Pairwise Weights in Practice . . . 58
3.6 Simulation Study . . . 59
3.6.1 Results . . . 60
3.7 Application to NHIS Data . . . 60
3.8 Discussion . . . 61

4 ESTIMATING THE EFFECT OF CLUSTER-LEVEL ADHERENCE ON AN INDIVIDUAL BINARY OUTCOME WITH A COMPLEX SAMPLING DESIGN . . . 63

4.1 Outline . . . 63
4.2 Introduction . . . 63
4.2.1 Instrumental Variable . . . 64
4.2.2 Structural Nested Models . . . 66
4.3 Methods . . . 68
4.3.1 Estimation with complex sampling designs . . . 68
4.3.2 Variance Estimation . . . 70
4.4 Simulation Study . . . 71
4.4.1 Simulation with strong instrument . . . 71
4.4.2 Simulation with weak instrument . . . 74
4.5 Results on School-based Water, Sanitation, and Hygiene (SWASH) Project 75
4.6 Conclusions . . . 77

APPENDIX

A PROOF OF LEMMA 1 (CHAPTER 2) . . . 78

B PROOF OF LEMMA 2 (CHAPTER 2) . . . 80

C SAS CODE (CHAPTER 3) . . . 81

D ESTIMATION OF INSTRUMENTAL VARIABLE ON STRUCTURE NESTED MODELS . . . 83

E VARIANCE ESTIMATION . . . 88

REFERENCES . . . 105

BIOGRAPHICAL SKETCH . . . 110

LIST OF TABLES

Table page

1-1 Results of the moderately biased sampling simulation 28

1-2 Results of the strongly biased sampling simulation 29

1-3 Results of the unbiased sampling simulation 31

1-4 Results of analyzing the NHIS data 32

2-1 Results of simulation study with $\boldsymbol{\beta} = (0.5, 0.8)$. The MLE is $\hat{\boldsymbol{\beta}} = (1.0814, 1.6508)$ with $SD(\hat{\boldsymbol{\beta}}) = (0.3838, 0.1935)$. 48

4-1 Distribution of $Y(0)|A = a, Z$. 72

4-2 Distribution of $Y(a)|A = a, Z$. 73

4-3 Simulation results for IV method on a structural nested model with strong instrument. 73

4-4 Simulation results for IV method on a structural nested model with weak instrument. 75

LIST OF FIGURES

Figure page

2-1 $f_k(\beta)$ as a function of k for $\beta = 0.5$ 45

2-2 $f_k(\beta)$ as a function of k for $\beta = -0.5$ 46

4-1 DAG represents the causal relationship between variables. 66

4-2 DAG represents the causal relationship between variables (with individual level confounders). 68

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

CAUSAL INFERENCE WITH COMPLEX SAMPLING DESIGNS

By

Zhulin He

August 2012

Chair: Babette Brumback
Major: Biostatistics

This PhD dissertation concerns causal inference with data from complex sampling designs. It consist of several aspects regarding newly developed or implemented methodologies with complex survey data, which are described as follows.

First, we consider the problem of adjusting for confounding by cluster in the context of complex multistage sampling and a binary outcome. We investigate three categories of approaches – ordinary logistic regression for survey data, with either no effect or a fixed effect for each cluster; conditional logistic regression extended for survey data; and generalized linear mixed model (GLMM) regression for survey data. We use theory, simulation, and analyses of the 2005 National Health Interview Survey (NHIS) data to compare and contrast all of these methods. One conclusion is that all of the methods perform poorly when the sampling bias is strong, which motivates us to find another method works properly with strong biased sampling designs.

We then show that for logistic regression with a simple match-pairs design, infinitely replicating observations and maximizing the conditional likelihood results in an estimator identical to the unconditional maximum likelihood estimator (MLE) with a fixed effect for each pair based on the original sample. Therefore, applying conditional likelihood methods to a pseudosample with observations replicated a large number of times can lead to a biased estimator. This casts doubt on one alternative approach to conditional logistic regression with complex survey data.

In the third chapter, we generalize binary conditional logistic regression for complex survey data by implementing the method based on a weighted pseudo-likelihood, in which the contribution from each neighborhood involves all pairs of cases and controls in the neighborhood. We show that it corresponds to an equivalent ordinary weighted log-likelihood formulation with binary outcomes. We explain how to program the method using standard software for ordinary logistic regression with complex survey data. We then apply the method to 2009 National Health Interview Survey (NHIS) public use data, to estimate the effect of education on health insurance coverage, adjusting for confounding by neighborhood.

Last, we concentrate on adjusting for unmeasured confounding of the effect of cluster-level adherence on an individual binary outcome with complex sampling designs. Seeking new methodologies for adjusting for confounding due to cluster effects, we use double inverse-probability weighting to adjust for the disproportionate sampling and the association of individual-level confounders with randomization. Then we develop and apply methods based on structural nested models to estimate effects of adherence assessed in terms of relative risk and risk difference, using cluster-level randomization as an instrumental variable and using the double weights to adjust for complex sampling and individual-level confounding. As an important application, we wish to estimate the effect of school-level adherence on individual absenteeism in the context of a school-based water, sanitation, and hygiene intervention (WASH) in Western Kenya.

CHAPTER 1
EFFORTS TO ADJUST FOR CONFOUNDING BY NEIGHBORHOOD USING COMPLEX SURVEY DATA

1.1 Outline

In this chapter, we focus our investigation on our study on the estimation of an individual exposure effect in the effect in the presence of confounding by neighborhood effects, motivated by an analysis of National Health Interview Survey (NHIS) public use data (See Brumback et al. [14]). In the analysis, we operationalize neighborhood as the secondary sampling unit of the survey, which consists of small groups of neighboring census blocks. Thus the neighborhoods may be sampled with unequal probabilities, as may be individuals within neighborhoods. We then develop and compare several approaches for the analysis of the effect of dichotomized individual-level education on the receipt of adequate mammography screening. We investigate three categories of approaches – ordinary logistic regression for complex survey data, with either no effect or a fixed effect for each cluster; a simple modification of conditional logistic regression for complex survey data; and generalized linear mixed model (GLMM) regression for complex survey data [51]. We use theory, simulation, and analysis of the NHIS data to compare and contrast all of these methods. One conclusion is that all of the methods perform poorly when the sampling bias is strong, which motivates us to find another method that works properly with strongly biased sampling designs.

1.2 Introduction

In social epidemiology, the population may be widely distributed geographically. Social epidemiologists often consider neighborhood characteristics as some of the most important influences on health outcomes. Since individuals in the same neighborhood usually share similar characteristics (such as medical insurance coverage, health care access, standard of living), one may be interested in estimating the effects of the measured or unmeasured neighborhood characteristics, or estimating effects adjusted for confounding due to neighborhood characteristics. This research is motivated by

an analysis of National Health Interview Survey (NHIS) public use data [43] to assess the effect of individual level exposures on the receipt of a binary outcome: adequate mammography screening (mammogram within past 24 months) in women above 40. In this chapter, we would like to investigate different methods of adjusting for confounding by neighborhood using complex survey data with possibly informative sampling designs.

Recently, the problem of adjusting for confounding in causal inference has generated a large literature. Greenland, Robins, and Pearl [25] provided a historical review of the concept of confounding and discussed the distinction between confounding and collapsibility. They pointed out that confounding is one of many problems that plague studies of cause and effect. Greenland and Robins [24], Greenland and Morgenstern[23], Pearl [48], Hernán, Brumback, and Robins [29] discussed the problems in control of confounding. Some recent papers study the problem of adjusting for confounding by cluster. Berlin et al. [8] presented an example of adjusting confounding due to cluster (centers) in a clinical study. Localio, Berlin, and Ten Have [40] investigated the performance of several regression methods – conditional regression, random effects models, generalized estimating equations (GEE), survey sampling methods, and fixed effects regression – for adjusting for confounding by cluster (center) using simulations. They also pointed out and verified that for a large number of cluster size and small sample sizes within cluster, the fixed effects estimator is positively biased.

Unfortunately, in our literature search, we found no literature that explored the topic of adjusting for confounding due to cluster with complex survey data. Agresti [1] described ordinary logistic regression and conditional logistic regression for simple cluster sampling designs, where ordinary logistic regression is conducted with a fixed effect for each cluster. However, similarly to Localio et al. [40], Agresti [1] explained that ordinary logistic regression leads to biased estimator of individual exposure effect when the number of sampled individuals within clusters stays fixed while the number of sampled clusters tends to infinity. Neyman and Scott [47] explained that this

problem is due to the number of nuisance parameters tending to infinity as the number of sampled clusters increases, which invalidates the theory of standard maximum likelihood estimation and inference. Generalized linear mixed models (GLMMs) with a logit link [1] are an alternative method. Neuhaus and McCulloch [46] showed that GLMMs fail to adequately adjust for confounding by cluster when the number of sampled clusters is small. Neuhaus and Kalbfleisch [44] added cluster average exposure into the model. Berlin et al. [8], Begg and Parides [7] applied this approach to some real data analyses. For some nonbinary outcomes, Verbeke, Spiessens and Lesaffre [63] developed conditional linear mixed models with longitudinal data. Goetgeluk and Vansteelandt [19] developed conditional generalized estimating equations (CGEE), comparing to generalized estimating equations (GEE) which was first introduced by Liang and Zeger [36]. But Goetgeluk and Vansteelandt [19] excluded the logit link in their paper.

In this chapter, we examine and investigate three categories of approaches – ordinary logistic regression, conditional logistic regression and GLMMs adjusting for confounding due to cluster on a binary outcome in the setting of complex survey data. In Section 1.3, we introduce the motivating NHIS data and describe the sampling design setting of the survey. Then we set up the modeling framework in Section 1.4 and explicate the three categories of approach in Section 1.5. In Section 1.6, we study the estimation performance on complex survey data with different sampling designs with respect to each approach. For ordinary logistic regression approach, we consider two design settings, either including or not including a fixed effect for each cluster. With regard to conditional logistic regression, we generalize the approach using a pseudo-sample method to replicate the observations according to integer-scaled approximations of the sampling weights. We then use GLMMs to adjust for confounding by cluster with simulated complex survey data. We base our GLMM program on the work of Rabe-Hesketh and Skrondal [51], using pseudo-likelihood methods. We then

compare the results of three categories of approaches with complex survey data in different design settings. In Section 1.7, we apply the methods to the NHIS data to estimate the effect of education on mammography screening, adjusting for confounding by neighborhood. Discursion comes to the last section.

1.3 Motivating NHIS Example

Our research is motivated by the analysis of 2005 NHIS public use data [43] to estimate the effect of individual level exposures on the receipt of adequate mammography screening (at least one mammogram within the last 2 years). In our study, we consider education, health insurance status, age category, and race/ethnicity as the individual level exposures, and measured census tract variables (such as poverty and segregation indices, and high-dimensional categorical variable that is neighborhood itself) as the neighborhood-level predictors. Because of unmeasured confounding of neighborhood, such as differential availability of health services, we want to adjust for confounding by neighborhood itself. Our concern is how could we adjust for confounding of individual level variables by unmeasured neighborhood factors using the high-dimensional categorical variable. We will focus on the effect of dichotomized education (more than high school versus high school or less) on the receipt of adequate mammography screening, adjusting for confounding by neighborhood. Since we wanted to measure neighborhood locally, we operationalized it as the secondary sampling unit (SSU) of the survey, which consists of small groups of neighboring census blocks.

The NHIS is an annual cross-sectional survey, using face-to-face interviews, that enlists a complex multistage sampling design, described in detail by Botman et al. [10]. It is the one of nation's largest in-person surveys, which is conducted annually by the National Center for Health Statistics regarding access to health care. In the survey, the U.S. is partitioned into approximately 2000 primary sampling units (PSUs), which are individual counties or groups of adjacent counties. Approximately 350 PSUs are selected into the sample with unequal sampling probability, and these same PSUs are

used for several years in a row. Then the PSUs are each subdivided into 20 substrata based on the census concentration of black and Hispanic persons. The substrata definitions are consistent across PSUs; hence in some PSUs, some of the substrata may be empty. SSUs are sampled with unequal probability from each substratum within each PSU varies. Then households are sampled within SSUs, and finally, one individual per household is sampled for the cancer component of the survey, on which the mammography questions are asked. We excluded men and women under 40 from our sample, therefore there are approximately 8300 women residing in approximately 4800 SSUs in the final sample; the number of sampled women per SSU ranged from 1 to 8.

1.4 Modeling Framework

Let $i = 1, \cdots, M$ index all neighborhoods in the population and $j = 1, \cdots, N_i$ index all individuals within neighborhood i. Let Y_{ij} be a binary outcome for individual j in neighborhood i, and X_{ij} be the corresponding individual-level covariate. Here in the NHIS study, Y_{ij} indicates the receipt of adequate mammography screening and X_{ij} denotes the education level. Then a model for the individual level exposure of X_{ij} on outcome Y_{ij} that accounts for confounding by unmeasured neighborhood effects b_i is

$$h(E(Y_{ij}|X_{ij}, b_i)) = X_{ij}\beta + b_i, \tag{1–1}$$

where h is the logit link function. Let $X_i = (X_{i1}, \cdots, X_{iN_i})$ and $Y_i = (Y_{i1}, \cdots, Y_{iN_i})$. We have three assumptions

(1) $Y_{ij} \coprod \{X_{ij'}\}_{j' \neq j} | b_i$, which means X_{ij} is the only component of X_i that influences the distribution of Y_{ij} given b_i;

(2) $(Y_i, X_i, b_i) \coprod (Y_{i'}, X_{i'}, b_{i'})$ for $i \neq i'$, which indicates that individuals from different neighborhoods are independent;

(3) $Y_{ij} \coprod Y_{ij'} | X_i, b_i$ for $j \neq j'$, which means the individual outcomes are independent of each other given all information with respect to neighborhood i, including all individual exposures and neighborhood effects.

Let Y_{ij}^x be the potential outcome regarding $X_{ij} = x$. When the neighborhood is the only confounder, we assume that

$$\{Y_{ij}^x\} \coprod X_{ij} | b_i, \tag{1–2}$$

where $\{Y_{ij}^x\}$ denotes the set of potential outcomes defined by all possible values of x. Besides, we also need a consistency assumption, which assume that the potential outcomes Y_{ij}^x are well defined and that those for a given individual independent of the realized exposures for other individuals. Then a causal model can be set up based on the potential outcomes $\{Y_{ij}^x\}$ regarding all possible values of x

$$h(E(Y_{ij}^x | b_i)) = x\beta + b_i, \tag{1–3}$$

where β in model (1–1) is identical to β in the model (1–3).

However, in some conditions, the consistency assumption and causal model (1–3) could be dubious. For example, regarding the NHIS study, if X_{ij} indicates individual education level and b_i represented a neighborhood mean education level (that is, b_i is inextricably linked with $X_i = (X_{i1}, \cdots, X_{iN_i})$), then we can not manipulate X_{ij} without in turn manipulating b_i. In this case, model (1–3) would be nonsensical and the consistency assumption would be violated too. Please refer Greenland [22] for more related commentary. Instead, one might prefer to assume that b_i is only linked to stationary characteristics of the neighborhood, such as availability of health services, to avoid this violation.

In this chapter, we prefer model (1–1) rather than model (1–3). This is because model (1–1) will not be violated even when a causal interpretation is untenable, such as the cases presented in the last paragraph; one may also want to know the association

level between individual exposures and outcomes can be accounted for by unmeasured neighborhood-level factors.

1.5 Approaches to Estimation with Complex Survey Data

In this section, we elaborate the estimation approaches, ordinary logistic regression, conditional logistic regression and GLMM with complex survey data respectively. Suppose the number of sampled neighborhood is m and the number of sampled individual within neighborhood i is n_i. Let p_i denote the probability of selecting neighborhood i (corresponding to a SSU), and let $p_{j|i}$ denote the probability of selecting individual j within neighborhood i, conditional on having selected neighborhood i. Then it is easy to see that the unconditional probability of selecting individual j within neighborhood i is $p_{ji} = p_i p_{j|i}$. Correspondingly, define the weights w_i, $w_{j|i}$ and w_{ji} as the inverse probability, i.e. $1/p_i$, $1/p_{j|i}$ and $1/p_{ji}$.

We use the method of variance estimation based on the so-called design-consistent ultimate clusters [58]. As common in complex survey data analysis, we approximate the sampling distribution of the estimators as arising from resampling PSUs with replacement within primary strata, ignoring subsequent steps of the multistage design. Especially for the conditional logistic regression method, we approximate the sampling distribution as arising from resampling PSUs with replacement ignoring the primary strata. Suppose population of M neighborhoods and N_i individuals per neighborhood in model (1–1) is conceptualized as a random two-stage sample from the superpopulation [58]. Then model (1–1) holds except the condition that M and N_i approaches infinity.

In the next subsections, we will present the three estimation approaches, ordinary logistic regression, conditional logistic regression and GLMM with complex survey data respectively.

1.5.1 Ordinary Logistic Regression for Complex Survey Data

In the past decade, ordinary logistic regression and conditional logistic regression methods have been commonly used in survey data analysis, because many variables

are generally measured as categorical and the two methods can incorporate a large number of explanatory variables [33]. For ordinary logistic regression with complex survey data, we can use SAS PROC SURVEYLOGISTIC to estimate the effect of X_{ij} on the outcomes. with this approach, we would simply regression Y_{ij} on the intercept and X_{ij} using weighted logistic regression (with weights w_{ij}) and adjust the standard errors for the multistage design using Taylor-series linearization (See [34], [58]).

Another option for ordinary logistic regression is that setting the b_i as fixed effects in the logistic regression. However, this approach leads to inconsistent estimators of β due to the Neyman-Scott problem [47], i.e. when the number of sampled individuals per cluster stays fixed, the number of sampled clusters tends to infinity. In our NHIS study, n_i is fairly small (ranging from 1 to 8), and the number of clusters, m, comparing to n_i is relatively large. Therefore we would not consider this approach to lead to a good estimation for our study.

1.5.2 Conditional Logistic Regression for Complex Survey Data

If the clusters in one survey data were sampled with equal probability and individuals within cluster were also sampled with equal probability, conditional logistic regression has an advantage for adjusting for confounding by cluster. One can implement this procedure using SAS PROC LOGISTIC with the stratum statement. The conditional logistic regression is based on sufficient statistics for the b_i, which are $\sum_{j=1}^{N_i} Y_{ij}$, given the cluster and individuals within cluster are sampled with equal probability from superpopulation. Then we can derive the maximum likelihood estimator of the conditional likelihood. Agresti [1] described properties of conditional logistic regression in his book. Unfortunately, in our literature search, there was no published research regarding the generalization of conditional logistic regression for use with complex survey data.

For complex survey data with no equal restriction on sampling probabilities, denote the sufficient statistic for b_i as S_i; then, the conditional likelihood for the population is of

the form

$$
\begin{aligned}
L(\beta|S_i, i=1,\cdots,M) &= \prod_{i=1}^{M}\prod_{j=1}^{N_i} P(Y_{ij}|X_i, b_i, S_i) \\
&= \prod_{i=1}^{M} P(Y_{i1},\cdots,Y_{iN_i}|X_i, b_i)/P(S_i|X_i, b_i) \\
&= \prod_{i=1}^{M} \frac{\left(\prod_{j=1}^{N_i}\exp(X_{ij}\beta)^{Y_{ij}}\right)}{\left(\sum_{Y_i^*\in C_i}\prod_{j=1}^{N_i}\exp(X_{ij}\beta)^{Y_{ij}^*}\right)} \qquad (1\text{–}4)
\end{aligned}
$$

where $S_i = \sum_{j=1}^{N_i} Y_{ij}$ and C_i denote the set of all possible N_i-vectors of binary elements $Y_i^* = (Y_{i1}^*,\cdots,Y_{iN_i}^*)$ such that $\sum_{j=1}^{N_i} Y_{ij}^* = S_i$. Due to complicated presence of the denominator $P(S_i|X_i, b_i)$ in (1–4), it seems impossible to find a consistent estimate of the population conditional likelihood using pseudo-likelihood with weights w_i and $w_{j|i}$. We turn to integer-scaled weights instead of the inverse probability weights w_i and $w_{j|i}$ to construct a pseudosample with pseudo-likelihood. Specifically, we scale w_i and $w_{j|i}$ to be integer-valued, and denote the new integer-scaled weights as w_i^* and $w_{j|i}^*$. In practice, one can scale the weights and convert them to integers by dividing all of the weights by the smallest weight, multiplying by a moderate number, and then rounding all of the weights to the nearest integer. The moderate number needs to be small enough so that the estimation procedure concludes in a reasonable amount of time and without exceeding memory limitations. Then the pseudosample is constructed by replicating observations according to the weights, i.e. replicating individual j within neighborhood i $w_{j|i}^*$ times and then replicating neighborhood i w_i^* times. Then, the sampling distribution of the estimator could be approximated using the bootstrap that resampling the PSUs with replacement ignoring primary strata.

This approach may work better than ignoring the complex sampling design and applying conditional logistic regression to the unweighted data. However the approach fails to lead an unbiased estimator when the weights $w_{j|i}^* = k$ tends to be large. He and Brumback [28] showed that when k approaches to infinity based on the psuedosample,

the conditional maximum likelihood estimator is equivalent to the ordinary maximum likelihood estimator. A proof and details of the equivalence will be provided in Chapter 2. Due to Neyman-Scott problem [47], the estimator of the conditional logistic regression would be biased when it approximates the ordinary logistic regression estimator with a fixed effect for each cluster.

1.5.3 GLMM Regression for Complex Survey Data

There are two standard solutions to the Neyman-Scott problem [47]. One is conditional logistic regression in the context of model (1–1) for simple cluster sampling (substituting m for M and n_i for N_i). The other is the GLMM regression, specifically is logistic regression with random intercepts (one per cluster). To implement the GLMM regression, we need to specify a distribution of the b_i. The standard and most commonly used specification is i.i.d. $N(0, \delta_b^2)$. Neuhaus, Kalbfleisch and Hauck [45] stated one disadvantage of the GLMM that the choice of distribution for the b_i can effect the consistency of the GLMM estimator of β. Regarding our research on complex survey data with confounding by neighborhood effect, another disadvantage of the GLMM is that it presumes that b_i is independent of X_{ij}, while it is clearly violated when there is confounding by neighborhood.

To describe the second weakness of the GLMM, we assume model (1–1) holds for the population, we add another model of the form with respect to b_i

$$b_i = E(b_i|X_i) + \delta_i', \tag{1–5}$$

where $X_i = (X_{i1}, \cdots, X_{iN_i})$, δ_i' follows i.i.d. $N(0, \tau^2)$, and δ_i' is assumed to be independent of X_i. Moreover we assume that

$$E(b_i|X_i) = q(X_i; \psi), \tag{1–6}$$

where $q(X_i; \psi)$ is a linear function of ψ and ψ is finite-dimensional. It follows that

$$E(Y_{ij}|X_i, \delta_i') = h^{-1}(X_{ij}\beta + q(X_i; \psi) + \delta_i'), \tag{1–7}$$

where the distribution of the δ_i' conditional on X_i is i.i.d. $N(0, \tau^2)$. Furthermore, the β of (1–7) has the same interpretation as the β in model (1–1) due to the assumption that model (1–1) holds. Therefore, we have constructed a GLMM for the population with random effect δ_i' independent of X_{ij}; hence it is no longer matters that b_i is associated with X_{ij}. The approach of Neuhaus and Kalbfleisch [44] can be derived as a special case of model (1–5), where $\psi = (\psi_0, \psi_1)^T$, $q(X_i; \psi) = \psi_0 + \psi_1 \bar{X}_i$ and $\bar{X}_i = (1/N_i) \sum_{j=1}^{N_i} X_{ij}$. This leads to the population GLMM

$$E(Y_{ij}|X_i, \delta_i) = \text{expit}(\psi_0 + \psi_1 \bar{X}_i + X_{ij}\beta + \tau\delta_i), \tag{1–8}$$

where $\text{expit}(x) = \exp(x)/(1 + \exp(x))$, with $\delta_i = \delta_i'/\tau$ having a standard normal distribution given X_i as i.i.d. $N(0, 1)$. Following Grilli and Pratesi [26], we define $\theta = (\psi_0, \psi_1, \beta, \tau)^T$. Then we write the logarithm of the likelihood corresponding to (1–8) as

$$\log L(\theta) = \sum_{i=1}^{M} \log \int_{-\infty}^{\infty} \left[\exp\{\sum_{j=1}^{N_i} \log L_{ij}(\theta; \delta)\} \right] \phi(\delta) d\delta. \tag{1–9}$$

Assume that the number of sampled clusters and the number of sampled individuals within clusters tend to infinity, then we have pseudo-likelihood with respect to population likelihood (1–9)

$$\log \hat{L}(\theta) = \sum_{i=1}^{m} w_i \log \int_{-\infty}^{\infty} \left[\exp\{\sum_{j=1}^{n_i} w_{j|i} \log L_{ij}^{w}(\theta; \delta)\} \right] \phi(\delta) d\delta, \tag{1–10}$$

where $L_{ij}^{w}(\theta; \delta)$ is identical to $L_{ij}(\theta; \delta)$ but with $\bar{X}_i^W = \sum_{j=1}^{n_i} w_{j|i} X_{ij} / \sum_{j=1}^{ni} w_{j|i}$ substituted for $\bar{X}_i$. As explained in Grilli and Pratesi [26] for the special case of $\psi_1 = 0$, and in Pfeffermann et al. [49] for another case of the identity link function and $\psi_1 = 0$, it shows that the maximum pseudo-likelihood estimator $\hat{\theta}_{MPL}$ approaches the finite population maximum likelihood estimator $\hat{\theta}$ as m and n_i approach M and N_i, respectively. Because of the consistency of $\hat{\theta}$ for the super-population parameter θ as M goes to infinity, we can speculate that $\hat{\theta}_{MPL}$ perform a good estimation as well.

One obstacle of applying maximum pseudo-likelihood estimators in the GLMM regarding NHIS study is that the cluster sizes in survey data are quite small. It is not obvious that the maximum pseudo-likelihood estimators $\hat{\theta}_{MPL}$ is consistent to the true values, since n_i is not close to N_i and the weights $w_{j|i}$ could be quite unbalanced within clusters. For finite population, the pseudo-likelihood results in a biased estimator since the pseudo-likelihood involves with $w_{j|i}$ and the measurement error in $\bar{X}_i^W$ as an estimator of $\bar{X}_i$. Regarding the term involving the $w_{j|i}$, there has been much discussion by statisticians about how to scale the $w_{j|i}$ weights so that for small cluster sizes the procedure has good performance (See [51], [26], [49]). One of the most intuitive suggestions is to scale the $w_{j|i}$ so that the sum of these weights over j equals n_i, the sample size for the clusters (See [49], [41]). However, there has been no discussion about we could find the implications of substituting $\bar{X}_i^W$ for $\bar{X}_i$, except that Grilli and Rampichini [27] discussed one solution regarding to this problem for the condition that when the $w_{j|i}$ are all equal to one and n_i is much less than N_i. Their results might lead to a suggestion that the consistency of the estimator of ψ_1 is affected, but the consistency of the estimator of β is not very much affected under the condition.

The GLLAMM software of Rabe-Hesketh and Skrondal [51] can be used to maximize the pseudo-likelihood (1–10) and to derive a 95% confidence interval for β. One might be interested in a standardized population-average effect for a hypothetical population in which the distribution of b_i is the same for each possible value of X_{ij}. Assume the distribution of b_i is its distribution in the entire population, and denote it as $F^*(b)$, then the estimator of standardized population-average effect is

$$\beta_p = \text{logit}\left(\int \text{logit}^{-1}(\beta + b)dF^*(b)\right) - \text{logit}\left(\int \text{logit}^{-1}(b)dF^*(b)\right), \qquad (1\text{–}11)$$

Then we could use the predicted random effects $\hat{\delta}_i$ and $\hat{\theta}_{MPL}$ to estimate the standardized population-average effect according (1–11)

$$\begin{aligned}\hat{\beta}_p = \quad & \text{logit}\left(\frac{w_i}{\sum_{i=1}^m w_i}\left[\frac{\sum_{j=1}^{n_i} w_{j|i}\text{expit}(\hat{\beta}+\hat{\psi}_0+\bar{X}_i^W\hat{\psi}_1)+\hat{\delta}_i}{\sum_{j=1}^{n_i} w_{j|i}}\right]\right) \\ & -\text{logit}\left(\frac{w_i}{\sum_{i=1}^m w_i}\left[\frac{\sum_{j=1}^{n_i} w_{j|i}\text{expit}(\hat{\psi}_0+\bar{X}_i^W\hat{\psi}_1)+\hat{\delta}_i}{\sum_{j=1}^{n_i} w_{j|i}}\right]\right).\end{aligned} \tag{1–12}$$

1.6 Simulation Study

As we stated in the previous sections, there is no method that performs perfectly for all conditions among the three categories of approaches. All of the proposed methods have limitations and it is not theoretically clear that which method perform better for a given application. Therefore we appeal to several simulations to further compare the methods.

We simulated complex survey data using (a) moderately biased sampling, (b) strongly biased sampling, and (c) unbiased sampling with respect to our estimand of interest. Set $M = m = 1000$ and $N_i = 1000$ for each cluster i. Let X_{ij} conditionally independent Bernoulli random variables with probability expit(μ_i), where μ_i are i.i.d. $N(0,1)$. Set $b_i = -5\bar{X}_i + \delta_i$, where δ_i are i.i.d. $N(0,1)$. Then we generated the binary outcome Y_{ij} according to model (1–1), where our target of estimation β is set equal to 0.5. We simulated the population with $M = 1000$ clusters. Next, we sampled from the population following the designs (a), (b) and (c) respectively. First, we defined a binary concordance variable C and set the values of C according the whether $Y_{ij} = X_{ij}$ or not. That is we divided the individual observations into concordant ($C = 1$) and discordant ($C = 0$) groups. Moreover, we generated for each observation an independent Bernoulli random variable B, where the distribution of B decides the level of sampling bias. For (a), we set the probability of the Bernoulli distribution of B equal to 0.6 if $C = 1$ and 0.4 if $C = 0$. For (b), we simply let $B = C$. For (c) unbiased sampling, we let B be an

i.i.d. Bernoulli variable with probability 0.5. Finally, we included observations into the sample with independent probability 0.002 if $B = 1$ and 0.004 if $B = 0$. We included $m = 1000$ neighborhoods in the sample, and n_i could be varied for each cluster due to our sampling scheme. According to the sampling scheme, sampled individuals were given weights $w_i = 1$, $w_{j|i} = 2$ if $B = 1$, and $w_{j|i} = 1$ if $B = 0$. Each simulation was repeated 100 times for each method of estimating β, where the true value is 0.5. We will describe how we conduct the estimation procedures using softwares in Section 1.6.1, Section 1.6.2 and Section 1.6.3, respectively. Then we will present the estimation results in Section 1.6.4.

1.6.1 Ordinary Logistic Regression for Complex Survey Data

We applied weighted ordinary logistic regression to the three categories of sampled data with and without including a fixed effect for each cluster. We used SAS PROC LOGISTIC (version 9.2) to conducted the approach of estimating. The data were simulated using SAS software.

1.6.2 Conditional Logistic Regression for Complex Survey Data

We applied conditional logistic regression to the simulated data based on a pseudo-sample. We used three different methods of constructing the pseudo-sample. First, we let the pseudo-sample be equal to the sample, i.e. all the weights equal to 1 (pseudosample 1). Second, we constructed the pseudo-sample with weights $w^*_{j|i} = 2$ for $B = 1$ and $w^*_{j|i} = 1$ for $B = 0$ (pseudosample 2). And third, we constructed the pseudo-sample with weights $w^*_{j|i} = 20$ for $B = 1$ and $w^*_{j|i} = 10$ for $B = 0$ (pseudosample 3). We can compare the effects of scaling weights of smaller values and larger values by comparing the estimators of the second and the third methods. We used SAS PROC LOGISTIC (version 9.2) with the stratum statement conducted the approach of estimating. The data were simulated using SAS software.

1.6.3 GLMM Regression for Complex Survey Data

We applied GLMM regression with and without weights to the simulated data. Here we scaled the weights so that the sum over j equals n_i as we mentioned in Section 1.5.3 (See [49], [41]). We treated the distribution of b_i as Gaussian, and tried fitting the naive model (1–1) and the population GLMM (1–8) (which adjusts for confounding) based on its pseudo-likelihood approximation (1–10). Besides, to explore the effects of the measurement error in $\bar{X}_i^W$ as an estimate of $\bar{X}_i$ we discussed before, we also fit model (1–8) based on the pseudo-likelihood approximation (1–10) but without the measurement error. The last approach would be impossible in practice unless $\bar{X}_i$ were observed, but it provides us with useful information on the effects of measurement error in this simulation study. The GLLAMM macro with 20 adaptive quadrature points and the pweight option in Stata version 10 was used for the estimation. The data were simulated using Stata.

1.6.4 Results

The results for the moderately biased sampling scheme are presented in Table 1-1. Comparing the relative biases of all eight methods, it is obvious that conditional logistic regression with the smaller pseudo-sample performs the best (method 4, mean $\hat{\beta} = 0.46$; true $\beta = 0.5$). Either ordinary logistic regression without cluster effect (method 1, mean $\hat{\beta} = -0.29$) or naive GLMM approach assuming $\psi_1 = 0$ without accounting for confounding (method 6, mean $\hat{\beta} = -0.20$) shows disastrous results due to the incorrect estimated direction of the effect. The other two GLMM methods that correctly adjust for confounding (method 7, mean $\hat{\beta} = 0.36$ and method 8, mean $\hat{\beta} = 0.40$) represent similar and nearly correct results. Ordinary logistic regression with cluster effect (method 2, mean $\hat{\beta} = 0.62$) leads to a better results comparing to method 1. Comparing method 4 to conditional logistic regression with larger scaled weights (method 5, mean $\hat{\beta} = 0.59$), it indicates that method 5 uses excessive replication and the scale of the weights indeed

has impacts on the estimation procedure. The effect of measurement error in GLMM procedure is not obvious comparing the results of methods 6 and 7.

The results for the strongly biased sampling scheme are presented in Table 1-2. All of the eight methods perform poorly on estimating the true β. Failing to adjust the cluster effect in ordinary logistic regression (method 1, mean $\hat{\beta} = -0.29$), method 3 (mean $\hat{\beta} = -0.92$), all GLMM methods (method 6, mean $\hat{\beta} = -0.80$; method 7, mean $\hat{\beta} = -0.27$ and method 8, mean $\hat{\beta} = -0.14$) indicate the incorrect estimation direction. The results of methods 2 (mean $\hat{\beta} = 0.14$), method 4 (mean $\hat{\beta} = 0.10$), and method 5 (mean $\hat{\beta} = 0.12$) are quite similar, indicating that the Neyman-Scott problem [47] is either not a problem for this simulation setting. Comparing method 3, method 4 and method 5, we clearly observe that failing to weight the sample at all can lead to much bias in the conditional logistic regression, and the excessive weights have unconspicuous effects on the estimation procedure. We also observe that the conditional logistic regression methods that incorporate the weighting perform much better than the all three GLMM methods. Comparing method 7 and method 8, we find that including measurement error does lead to additional bias, which could be owing to poor performance of the GLMM methods. To further investigate the problem of the failure of the GLMM methods, we conducted a simulation without the confounding; i.e. we set $b_i = -0.5 + \delta_i$, so that b_i is independent of X_i. Then, we used the 'naive' method which is the best GLMM method among the eight methods for this new simulation setting. It turned out that the mean estimate of β is -0.18 with a standard deviation of 0.11 in 100 iterations of the simulation. It shows disastrous performance while the true value is 0.5. This implies that the methodology of Rabe-Hesketh and Skrondal [51] does not work well when the sampling bias is strong.

The results for the unbiased sampling are presented in Table 1-3. Similar to Table 1-1 and Table 1-2, we find methods 1 (mean $\hat{\beta} = -0.28$) and method 6 (mean $\hat{\beta} = -0.04$) fails to pointed out the correct direction of estimation owing to failing to

Table 1-1. Results of the moderately biased sampling simulation

Category	Method	Mean	SD	Range	Relative bias
Ordinary logistic regression					
	(1) without cluster effect	-0.29	0.12	(-0.61, -0.05)	1.58
	(2) With cluster effect	0.62	0.22	(-0.01, 1.11)	0.24
Conditional logistic regression					
	(3) Pseudosample 1	0.21	0.16	(-0.22, 0.78)	0.58
	(4) Pseudosample 2	0.46	0.18	(-0.05, 1.05)	0.08
	(5) Pseudosample 3	0.59	0.20	(-0.04, 1.05)	0.18
GLMM					
	(6) Naive	-0.20	0.13	(-0.45, 0.06)	1.40
	(7) Adjusted	0.36	0.16	(0.04, 0.80)	0.28
	(8) Adjusted without measurement error	0.40	0.13	(0.06, 0.68)	0.20

Mean, SD, and range are based on 100 simulations. Different simulated data sets are used for each result, except for the pairs (pseudosample 1, pseudosample 2) and (naive, adjusted). True value of estimand is 0.50. Relative bias is calculated as the absolute value of $(\text{mean} - 0.5)/0.5$.

Table 1-2. Results of the strongly biased sampling simulation

Category	Method	Mean	SD	Range	Relative bias
Ordinary logistic regression					
	(1) without cluster effect	-0.29	0.11	(-0.61, 0.11)	1.58
	(2) With cluster effect	0.14	0.22	(-0.40, 0.70)	0.72
Conditional logistic regression					
	(3) Pseudosample 1	0.92	0.14	(-1.19, -0.58)	2.84
	(4) Pseudosample 2	0.10	0.15	(-0.22, 0.47)	0.80
	(5) Pseudosample 3	0.12	0.22	(-0.36, 0.86)	0.76
GLMM					
	(6) Naive	-0.80	0.15	(-1.13, 0.44)	2.60
	(7) Adjusted	-0.27	0.18	(-0.62, 0.16)	1.54
	(8) Adjusted without measurement error	-0.14	0.16	(-0.53, 0.19)	1.24

Mean, SD, and range are based on 100 simulations. Different simulated data sets are used for each result, except for the pairs (pseudosample 1, pseudosample 2) and (naive, adjusted). True value of estimand is 0.50. Relative bias is calculated as the absolute value of $(\text{mean} - 0.5)/0.5$.

adjust or to properly adjust for the neighborhood effect. Besides, the failure of method 6 emphasizes the problem with standard applications of GLMMs to adjust for confounding by neighborhood, where these methods all assume that $E(b_i|X_i) = \psi_0$. Comparing to method 3 (no replications; mean $\hat{\beta} = 0.51$), method 4 (few replications; mean $\hat{\beta} = 0.59$), and method 5 (excessive replications; mean $\hat{\beta} = 0.71$), it shows the adverse effect of replicating observations on the conditional logistic regression, which means no weighting technology is needed in the conditional logistic regression. Finally, comparing methods 7 (mean $\hat{\beta} = 0.54$) and method 8 (mean $\hat{\beta} = 0.53$), we reach a conclusion that for the unbiased sampling scheme, the effect of measurement error on the performance of the GLMM methods is negligible.

1.7 Application to NHIS Data

Table 1-4 presents the results of applying the methods of Section 1.5 to the NHIS data to estimate the effect of education on receipt of adequate mammography screening, adjusting for confounding by neighborhood. Estimated odds ratios and 95% confidence intervals are presented for five methods: ordinary weighted logistic regression without a cluster effect, conditional logistic regression with replication, GLMM with the naive model $E(b_i|X_i) = \psi_0$, GLMM with the model for confounding adjustment, $E(b_i|X_i) = \psi_0 + \psi_1 \bar{X}_i^W$, and the GLMM estimator of the population-averaged effect, β_p, computed according to equation (1–12).

All of the five estimators shows the same direction of the education effects on the outcome. We see that the ordinary logistic regression without cluster effect (method 1, $OR = 1.62$, 95% CI (1.45, 1.81)) and the GLMM regression with naive model (method 3, $OR = 1.66$, 95% CI (1.48, 1.86)) are similar. Both conditional logistic regression with replication (method 2, $OR = 1.44$, 95% CI (1.15, 1.79)) and the adjusted GLMM regression (method 4, $OR = 1.24$, 95% CI (1.05, 1.46)) show the reduction of the effect of education after adjusting for confounding by neighborhood, which is reasonable. The GLMM estimate of the population-averaged effect (method 5, $OR = 1.18$, 95% CI (1.04,

Table 1-3. Results of the unbiased sampling simulation

Category	Method	Mean	SD	Range	Relative bias
Ordinary logistic regression					
	(1) without cluster effect	-0.28	0.12	(-0.54, 0.14)	1.56
	(2) With cluster effect	0.66	0.21	(-0.07, 1.15)	0.32
Conditional logistic regression					
	(3) Pseudosample 1	0.51	0.16	(0.13, 0.99)	0.02
	(4) Pseudosample 2	0.59	0.19	(0.19, 1.12)	0.18
	(5) Pseudosample 3	0.71	0.22	(0.23, 1.19)	0.42
GLMM					
	(6) Naive	-0.04	0.13	(-0.33, 0.35)	1.08
	(7) Adjusted	0.54	0.16	(0.21, 0.91)	0.08
	(8) Adjusted without measurement error	0.53	0.13	(0.27, 0.84)	0.06

Mean, SD, and range are based on 100 simulations. Different simulated data sets are used for each result, except for the pairs (pseudosample 1, pseudosample 2) and (naive, adjusted). True value of estimand is 0.50. Relative bias is calculated as the absolute value of $(\text{mean} - 0.5)/0.5$.

1.35)) is quite different from the estimate of the population-averaged effect in method 1, which does not adjust for confounding by neighborhood.

Table 1-4. Results of analyzing the NHIS data

Method	OR estimate	95% CI
(1) Ordinary logistic regression without cluster effect	1.62	(1.45, 1.81)
(2) Conditional logistic regression with replication	1.44	(1.15, 1.79)
(3) GLMM naive	1.66	(1.48, 1.86)
(4) GLMM adjusted	1.24	(1.05, 1.46)
(5) GLMM population-averaged effect, adjusted	1.18	(1.04, 1.35)

1.8 Discussion

We concentrate our study on the problem of adjusting for confounding by cluster in the context of complex survey data with binary outcome. We have investigated three categories of approaches – ordinary logistic regression for complex survey data, with either no effect or a fixed effect for each cluster; conditional logistic regression extended for complex survey data; and generalized linear mixed model (GLMM) regression for complex survey data – in terms of theory, simulation, and practice. The theory indicates that all of the methods considered have limitations. Our simulation study indicates that all of the methods perform poorly when the sampling bias is strong and some of the methods lead to a good estimator when there is no sampling bias or moderately sampling bias. This motivates us to find another method works properly with strong biased sampling designs.

CHAPTER 2
AN EQUIVALENCE OF CONDITIONAL AND UNCONDITIONAL MAXIMUM LIKELIHOOD ESTIMATORS VIA INFINITE REPLICATION OF OBSERVATIONS

2.1 Outline

In this chapter, we prove that for logistic regression with a match-pairs design with binary outcome, after infinitely replicating the observations, the conditional maximum likelihood estimator (CMLE) is equivalent to the unconditional maximum likelihood estimator (MLE) with a fixed effect for each pair based on the original sample. We also show the CMLE for the pseudosample that replicating the original sample k time falls in a range built by MLE for the original sample. It is known that for binary matched-pairs data with single predictor, the unconditional logistic regression estimator is biased. Therefore, applying conditional likelihood methods to a pseudosample with observations replicated a large number of times can lead to a biased estimator. This result casts doubt on one possible approach to conditional logistic regression with complex survey data.

2.2 Introduction

In Chapter 1, motivated by an application with complex survey data, we compared several methods, including conditional logistic regression and unconditional logistic regression, with a simulated complex survey data [14]. The results show that for conditional logistic regression, replicating observations more times (i.e. larger weights) would result different estimator to the estimator that replicating observations less times (i.e. smaller weights). In addition, as the number of times that observations were replicated increases to infinity, we speculate an asymptotic equivalence between the conditional logistic regression and unconditional logistic regression methods. Therefore, applying conditional likelihood methods to a pseudosample with observations replicated a large number of times can lead to a biased estimator. One consequence is that a possible approach to applying conditional likelihood methods to complex survey data is prone to bias. Specifically, applying conditional likelihood methods to a pseudosample

constructed by replicating observations according to integer-scaled versions of complex sampling weights can lead to a biased estimator when the unconditional maximum likelihood methods are biased.

In the logistic regression context, our modelling framework is as follows. Let i index clusters and let j index individuals within cluster in the population. To estimate the effect of a p-dimensional individual-level exposure X_{ij} and a binary outcome Y_{ij}, one population-level model that adjusts for confounding by cluster is given by

$$\text{logit}(E(Y_{ij}|X_{ij}, b_i)) = X_{ij}^T\beta + b_i, \tag{2–1}$$

where β is the effect of interest and all outcomes are assumed independent of one another given the cluster effects, b_i. With ordinary cluster sampling, m clusters and n_i individuals within each cluster i are selected at random from a conceptually infinite population of clusters and individuals within cluster. Neyman and Scott [47] pointed out that the asymptotic theory for unconditional maximum likelihood estimation is invalid for estimating β and the collection of b_i when the n_i remain small while m approaches infinity. Andersen [2] and Agresti [1] points out that for one-dimensional covariate model of 2–1, the unconditional maximum likelihood estimator (MLE) of β is inconsistent in that setting, and especially so for the simple matched pairs design in which $n_i = 2$, $X_{i1} = 0$ and $X_{i2} = 1$. For that design, let n_{12} be the number of observations with $Y_{i1} = 1$ and $Y_{i2} = 0$, and n_{21} be the number of observations with $Y_{i1} = 0$ and $Y_{i2} = 1$. Then the MLE of β is $2\log(n_{21}/n_{12})$, whereas the conditional maximum likelihood estimator (CMLE), which is consistent, is $\log(n_{21}/n_{12})$. Thus, the MLE is biased by a factor of two for the simple matched pairs design.

For more general designs enlisting ordinary cluster sampling, denote the contribution to the likelihood corresponding to (2–1) from individual j of cluster i as $L_{ij}(\beta, b_i) = \exp(X_{ij}^T\beta + b_i)^{Y_{ij}}/(1 + \exp(X_{ij}^T\beta + b_i))$, and let $L_i(\beta, b_i) = \prod_{j=1}^{n_i} L_{ij}(\beta, b_i)$, $b = (b_1, \ldots, b_m)$, and $L(\beta, b) = \prod_{i=1}^{m} L_i(\beta, b_i)$. Let $\theta = (\beta, b)$ and let $l(\theta)$ and $l_{ij}(\theta)$ denote the natural

logarithms of $L(\beta, b)$ and $L_{ij}(\beta, b_i)$, respectively. The unconditional maximum likelihood estimator of θ solves $\frac{\partial l(\theta)}{\partial \theta} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{\partial l_{ij}(\theta)}{\partial \theta} = 0$. With complex survey data, denote the reciprocal of the probability of selecting cluster i into the sample as w_i and that of the conditional probability of selecting individual j from cluster i given selection of cluster i as $w_{j|i}$. Let $w_{ij} = w_i w_{j|i}$. The maximum pseudolikelihood estimator (MPLE, [58]) for θ is defined as the solution to $\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\frac{\partial l_{ij}(\theta)}{\partial \theta} = 0$. If the weights w_{ij} were scaled to be integers, the MPLE would equal the MLE based on a pseudosample in which observation ij is replicated w_{ij} times. Unfortunately, the MPLE inherits inconsistencies from the MLE for settings in which w_{ij} is constant. Thus, it is not a method that will generally deliver consistent estimators of β.

Generalizing conditional maximum likelihood estimation for use with complex survey data is difficult (see [14], [20] and [15]), because the pseudolikelihood approach cannot be directly applied with a conditional likelihood. To understand why, denote the sufficient statistic for b_i as $S_i = \sum_{j=1}^{n_i} Y_{ij}$, and let V_i denote the set of all possible n_i-vectors of binary elements Y_i^* such that $\sum_{j=1}^{n_i} Y_{ij}^* = S_i$. Then, the conditional likelihood corresponding to (2–1) is of the form

$$\begin{aligned} L(\beta|S_i, i = 1, \dots, m) &= \prod_{i=1}^{m}\prod_{j=1}^{n_i} P(Y_{ij}|X_i, b_i, S_i) \\ &= \prod_{i=1}^{m} P(Y_{i1}, \dots, Y_{in_i}|X_i, b_i)/P(S_i|X_i, b_i) \\ &= \prod_{i=1}^{m} \frac{\prod_{j=1}^{n_i} \exp(X_{ij}^T\beta)^{Y_{ij}}}{\sum_{Y_i^* \in V_i}\prod_{j=1}^{n_i} \exp(X_{ij}^T\beta)^{Y_{ij}^*}}. \end{aligned} \tag{2–2}$$

The b_i drop out of $L(\beta|S_i, i = 1, \dots, m)$ due to conditioning on the S_i. Owing to the denominator term $P(S_i|X_i, b_i)$, the pseudolikelihood approach, in which contributions to the log-likelihood are weighted by the $w_{j|i}$ and w_i, is not possible. Nonetheless, it is possible to use integer-scaled weights to construct a pseudosample and then maximize

(2–2) for the pseudosample. Specifically, scale the weights w_{ij} to be integer-valued; denote these scaled weights as w_{ij}^*. In practice, one can scale the weights and convert them to integers by dividing all of the weights by the smallest weight, multiplying by a moderate number, and then rounding all of the weights to the nearest integer. The moderate number needs to be small enough so that the estimation procedure concludes in a reasonable amount of time and without exceeding memory limitations. Then, the pseudosample is constructed by replicating individual j within neighbourhood i w_{ij}^* times. The likelihood at (2–2) becomes

$$\frac{\prod_{j=1}^{n_i}\prod_{k=1}^{w_{ij}^*}\exp(X_{ij}^T\beta)^{Y_{ij}}}{\sum_{Y_i^*\in C_i^*}\prod_{j=1}^{n_i}\prod_{k=1}^{w_{ij}^*}\exp(X_{ij}^T\beta)^{Y_{ijk}^*}} \tag{2–3}$$

where C_i^* denotes the set of all possible binary vectors Y_i^* of size $\sum_{j=1}^{n_i} w_{ij}^*$ such that $\sum_{j=1}^{n_i}\sum_{k=1}^{w_{ij}^*} Y_{ijk}^* = \sum_{j=1}^{n_i}\sum_{k=1}^{w_{ij}^*} Y_{ij}$, and Y_{ijk}^* is one or zero depending on the designated element of C_i^*. We then maximize the product of (2–3) over i as a function of β to obtain the conditional likelihood estimator based on the pseudosample. The sampling distribution of the estimator can be approximated using a bootstrap that resamples PSUs (primary sampling units) with replacement ignoring primary strata; this is a conservative approach in that it can only overestimate the true sampling variability.

Intuition suggested, and simulation studies supported, the conjecture of Chapter 1 that the conditional maximum likelihood estimator based on a pseudosample with large-scale replication approximates the unconditional MLE based on that pseudosample. The intuition derives from the approximation of the conditional maximum likelihood estimator by a generalized linear mixed model estimator (e.g. see [37], [45], [55], [54] and [57]). For the generalized linear mixed model with a smooth mixing distribution having positive support on the entire real line, large-scale replication of observations would lead to an estimator that tends towards treating the random effects as fixed effects; treating the random effects as fixed effects leads to the unconditional MLE. In Section 2.3, we formalize and prove a result even stronger than the conjecture

for a special case, that of the matched pairs design. In Section 2.4, we conduct a simulation study in order to investigate a more complex scenario where there is no restriction on X_{ij}. Section 2.5 concludes with a discussion of implications and future directions.

2.3 Main Results

Our main result pertains to the matched pairs design of Section 2.2, in which $w_{j|i} = w_i = 1$.

Theorem. *For model (2–1) and the matched pairs design with $n_i = 2$, X_{i1} and X_{i2} are the p-dimensional covariates. Define $C_i = X_{i2} - X_{i1} = (C_{i1}, \cdots, C_{ip})_{p\times 1}$ and for at least one j where $C_j \neq 0$. Denote by $\hat{\beta}_c(k)$ the CMLE that maximizes (2–2) based on pretending that a pseudosample in which $w^*_{ij} = k$ is the observed sample, and $\hat{\beta}$ is the standard MLE based on the original sample. Then*

(1) As $k \to \infty$, $\hat{\beta}_c(k)$ approaches $\hat{\beta}$ monotonically; i.e. $\lim_{k\to\infty} \hat{\beta}_c(k) = \hat{\beta}$;

(2) $\hat{\beta}_c(1) = \frac{\hat{\beta}}{2}$.

Proof. We need the following two Lemmas.

Lemma 1. For $C_i \neq 0_{p\times 1}$,

$$\lim_{k\to\infty} \frac{\sum_{r=0}^{k} \binom{k}{r}^2 (k-r) e^{(k-r)C_i^T\beta}}{\sum_{r=0}^{k} \binom{k}{r}^2 k e^{(k-r)C_i^T\beta}} = \frac{e^{\frac{C_i^T\beta}{2}}}{1+e^{\frac{C_i^T\beta}{2}}}. \tag{2–4}$$

The proof of Lemma 1 is in Appendix A.

Lemma 2. *Without loss of generality, suppose $C_{is} > 0$, $s = 1, \cdots, p$. When $\beta_s \geq 0$, $\frac{\sum_{r=0}^{k} \binom{k}{r}^2 (k-r) e^{(k-r)C_{is}^T\beta_s}}{\sum_{r=0}^{k} \binom{k}{r}^2 k e^{(k-r)C_{is}^T\beta_s}}$ is non-increasing with $k \to \infty$; when $\beta_s < 0$, $\frac{\sum_{r=0}^{k} \binom{k}{r}^2 (k-r) e^{(k-r)C_{is}^T\beta_s}}{\sum_{r=0}^{k} \binom{k}{r}^2 k e^{(k-r)C_{is}^T\beta_s}}$ is non-decreasing with $k \to \infty$.*

The proof of Lemma 2 is in Appendix B.

Let $\mathbf{Y}^k_{ij} = (y_{ij1}, y_{ij2}, \cdots, y_{ijk})^T$, $j = 1, 2$, and $\mathbf{Y}^k_i = (\mathbf{Y}^{k\,T}_{i1}, \mathbf{Y}^{k\,T}_{i2})'$. Further let $S_i = \sum_j \sum_l y_{ijl}$. By construction of the pseudosample, $S_i = 0$, k or $2k$.

From the model (2–1), we have probabilities regarding two cases that $\{Y_{i1} = 1, Y_{i2} = 0\}$ and $\{Y_{i1} = 0, Y_{i2} = 1\}$,

$$P(Y_{i1} = 1, Y_{i2} = 0) = \frac{e^{b_i + x_{i1}^T \beta}}{1 + e^{b_i + x_{i1}^T \beta}} \frac{1}{1 + e^{b_i + x_{i2}^T \beta}},$$

and

$$P(Y_{i1} = 0, Y_{i2} = 1) = \frac{1}{1 + e^{b_i + x_{i1}^T \beta}} \frac{e^{b_i + x_{i2}^T \beta}}{1 + e^{b_i + x_{i2}^T \beta}}.$$

Supposing that model (2–1) generated the pseudosample, we have that

$$P(\mathbf{Y}_i^k = \mathbf{0} | S_i = 0) = 1,$$

and

$$P(\mathbf{Y}_i^k = \mathbf{1} | S_i = 2k) = 1.$$

Furthermore,

$$P(S_i = k) = \sum_{r=0}^{k} P((\sum_l y_{i1l}) = r \text{ and } (\sum_l y_{i2l}) = k - r),$$

so that

$$\begin{aligned} P(S_i = k) &= \sum_{r=0}^{k} \binom{k}{r} \left(\frac{e^{b_i + x_{i1}^T \beta}}{1 + e^{b_i + x_{i1}^T \beta}} \right)^r \left(\frac{1}{1 + e^{b_i + x_{i1}^T \beta}} \right)^{k-r} \binom{k}{r} \left(\frac{e^{b_i + x_{i2}^T \beta}}{1 + e^{b_i + x_{i2}^T \beta}} \right)^{k-r} \left(\frac{1}{1 + e^{b_i + x_{i2}^T \beta}} \right)^r \\ &= \sum_{r=0}^{k} \binom{k}{r}^2 \left(\frac{1}{1 + e^{b_i + x_{i1}^T \beta}} \right)^k \left(\frac{1}{1 + e^{b_i + x_{i2}^T \beta}} \right)^k \left(e^{b_i + x_{i1}^T \beta} \right)^r \left(e^{b_i + x_{i2}^T \beta} \right)^{k-r}. \end{aligned}$$

For the pseudosample, when $S_i = k$, either $\mathbf{Y}_{i1}^k = \mathbf{1}$ and $\mathbf{Y}_{i2}^k = \mathbf{0}$, or $\mathbf{Y}_{i1}^k = \mathbf{0}$ and $\mathbf{Y}_{i2}^k = \mathbf{1}$. Using the preceding expression,

$$\begin{aligned} P(\mathbf{Y}_{i1}^k = \mathbf{1}, \mathbf{Y}_{i2}^k = \mathbf{0} | S_i = k) &= \frac{\left(\frac{e^{b_i + x_{i1}^T\beta}}{1 + e^{b_i + x_{i1}^T\beta}}\right)^k \left(\frac{1}{1 + e^{b_i + x_{i2}^T\beta}}\right)^k}{\sum_{r=0}^k \binom{k}{r}^2 \left(\frac{1}{1 + e^{b_i + x_{i1}^T\beta}}\right)^k \left(\frac{1}{1 + e^{b_i + x_{i2}^T\beta}}\right)^k \left(e^{b_i + x_{i1}^T\beta}\right)^r \left(e^{b_i + x_{i2}^T\beta}\right)^{k-r}} \\ &= \frac{1}{\sum_{r=0}^k \binom{k}{r}^2 \left(e^{C_i^T\beta}\right)^{k-r}}, \end{aligned}$$

and

$$\begin{aligned} P(\mathbf{Y}_{i1}^k = \mathbf{0}, \mathbf{Y}_{i2}^k = \mathbf{1} | S_i = k) &= \frac{\left(\frac{1}{1 + e^{b_i + x_{i1}^T\beta}}\right)^k \left(\frac{e^{b_i + x_{i2}^T\beta}}{1 + e^{b_i + x_{i2}^T\beta}}\right)^k}{\sum_{r=0}^k \binom{k}{r}^2 \left(\frac{1}{1 + e^{b_i + x_{i1}^T\beta}}\right)^k \left(\frac{1}{1 + e^{b_i + x_{i2}^T\beta}}\right)^k \left(e^{b_i + x_{i1}^T\beta}\right)^r \left(e^{b_i + x_{i2}^T\beta}\right)^{k-r}} \\ &= \frac{\left(e^{C_i^T\beta}\right)^k}{\sum_{r=0}^k \binom{k}{r}^2 \left(e^{C_i^T\beta}\right)^{k-r}}. \end{aligned}$$

Note that $c_i = x_{i2} - x_{i1}$ and $I\{\mathbf{Y}_{i1}^k = \mathbf{1},\ \mathbf{Y}_{i2}^k = \mathbf{0}\}$ and $I\{\mathbf{Y}_{i1}^k = \mathbf{1},\ \mathbf{Y}_{i2}^k = \mathbf{0}\}$ indicate $\{\mathbf{Y}_{i1}^k = \mathbf{1},\ \mathbf{Y}_{i2}^k = \mathbf{0}\}$ and $\{\mathbf{Y}_{i1}^k = \mathbf{1},\ \mathbf{Y}_{i2}^k = \mathbf{0}\}$ respectively. The conditional likelihood (2–2) applied to the pseudosample is then

$$\begin{aligned} L_k(\beta | S_i, i = 1, \ldots, m) &= \prod_{i=1}^m \left[\frac{1}{\sum_{r=0}^k \binom{k}{r}^2 \left(e^{c_i^T\beta}\right)^{k-r}}\right]^{I\{Y_{i1}^k = 1,\ Y_{i2}^k = 0\}} \left[\frac{\left(e^{c_i^T\beta}\right)^k}{\sum_{r=0}^k \binom{k}{r}^2 \left(e^{c_i^T\beta}\right)^{k-r}}\right]^{I\{Y_{i1}^k = 0,\ Y_{i2}^k = 1\}} \\ &= \prod_{i=1}^m \frac{\left[\left(e^{c_i^T\beta}\right)^k\right]^{I\{Y_{i1}^k = 0,\ Y_{i2}^k = 1\}}}{\left[\sum_{r=0}^k \binom{k}{r}^2 \left(e^{c_i^T\beta}\right)^{k-r}\right]^{I\{Y_{i1}^k = 1,\ Y_{i2}^k = 0\} + I\{Y_{i1}^k = 0,\ Y_{i2}^k = 1\}}}. \end{aligned} \tag{2–5}$$

Thus the conditional log likelihood is

$$
\begin{aligned}
l_k(\beta) &= \sum_{i=1}^{m} k c_i^T \beta I\{\mathbf{Y}_{i1}^k = \mathbf{0},\ \mathbf{Y}_{i2}^k = \mathbf{1}\} \\
&\quad - \sum_{i=1}^{m} (I\{\mathbf{Y}_{i1}^k = \mathbf{1},\ \mathbf{Y}_{i2}^k = \mathbf{0}\} + I\{\mathbf{Y}_{i1}^k = \mathbf{0},\ \mathbf{Y}_{i2}^k = \mathbf{1}\}) \log \left[\sum_{r=0}^{k} \binom{k}{r}^2 \left(e^{(k-r)c_i^T \beta} \right) \right],
\end{aligned}
$$

which is maximized when

$$
\begin{aligned}
\frac{\partial l_k(\beta)}{\partial \beta_s} &= \sum_{i=1}^{m} k c_{is} I\{\mathbf{Y}_{i1}^k = \mathbf{0},\ \mathbf{Y}_{i2}^k = \mathbf{1}\} \\
&\quad - \sum_{i=1}^{m} c_{is} (I\{\mathbf{Y}_{i1}^k = \mathbf{1},\ \mathbf{Y}_{i2}^k = \mathbf{0}\} + I\{\mathbf{Y}_{i1}^k = \mathbf{0},\ \mathbf{Y}_{i2}^k = \mathbf{1}\}) \frac{\sum_{r=0}^{k} \binom{k}{r}^2 (k-r) e^{(k-r)c_i^T \beta}}{\sum_{r=0}^{k} \binom{k}{r}^2 e^{(k-r)c_i^T \beta}} = 0,
\end{aligned}
\tag{2–6}
$$

for all $s = 1, \cdots, p$.

Andersen [2] explained the idea how to estimate the MLE for simple matched pairs data. Now we use his method extend to p-dimensional covariates matched pairs designs. The model (2–1) can be rewritten as

$$
\text{logit}(E(Y_{ij} | X_{ij}^*, b_i)) = X_{ij}^{*T} \beta + b_i^*, \tag{2–7}
$$

where $X_{ij}^* = 0_{p \times 1}$ for $j = 1$ and $X_{ij}^* = C_i$ for $j = 2$ and $b_i^* = b_i + X_{i1}^T \beta$. Then the likelihood for the ith stratum is

$$
L_i = \pi_{i1}^{y_{i1}} (1 - \pi_{i1}^{1-y_{i1}}) \pi_{i2}^{y_{i2}} (1 - \pi_{i2}^{1-y_{i2}}),
$$

where $\pi_{i1} = \frac{\exp(b_i^*)}{1+\exp(b_i^*)}$, $\pi_{i2} = \frac{\exp(b_i^* + c_i^T \beta)}{1+\exp(b_i^* + c_i^T \beta)}$. The corresponding log-likelihood is

$$
l_i = y_{i1} \log \pi_{i1} + (1 - y_{i1}) \log(1 - \pi_{i1}) + y_{i2} \log \pi_{i2} + (1 - y_{i2}) \log(1 - \pi_{i2}).
$$

Then we have the partial differential equations as below

$$\begin{aligned}
\frac{\partial l_i}{\partial \pi_{i1}} &= \frac{y_{i1} - \pi_{i1}}{\pi_{i1}(1-\pi_{i1})}, \\
\frac{\partial l_i}{\partial \pi_{i2}} &= \frac{y_{i2} - \pi_{i2}}{\pi_{i2}(1-\pi_{i2})}, \\
\frac{\partial \pi_{i1}}{\partial \beta_s} &= 0, s = 1, \cdots, p, \\
\frac{\partial \pi_{i2}}{\partial \beta_s} &= c_{is}\pi_{i2}(1-\pi_{i2}), s = 1, \cdots, p.
\end{aligned}$$

Then the corresponding log likelihood can be expressed by

$$\begin{aligned}
l(\beta) &= \sum_{i=1}^{m} b_i^* I\{Y_{i1} = 1, Y_{i2} = 0\} + (b_i^* + c_i^T\beta) I\{Y_{i1} = 0, Y_{i2} = 1\} \\
&\quad + (2b_i^* + c_i^T\beta) I\{Y_{i1} = 1, Y_{i2} = 1\} - \log(1 + e^{b_i^*}) - \log(1 + e^{b_i^* + c_i^T\beta}), \quad (2\text{–}8)
\end{aligned}$$

which is maximized when

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta_s} &= \sum_{i=1}^{m} c_{is} \left(I\{Y_{i1} = 1, Y_{i2} = 1\} + I\{Y_{i1} = 0, Y_{i2} = 1\} \right) - \sum_{i=1}^{m} c_{is} \frac{e^{b_i^* + c_i^T\beta}}{1 + e^{b_i^* + c_i^T\beta}} \\
&= \sum_{i=1}^{m} c_{is} I\{Y_{i2} = 1\} - \sum_{i=1}^{m} c_{is} \frac{e^{b_i^* + c_i^T\beta}}{1 + e^{b_i^* + c_i^T\beta}} \\
&= \sum_{i=1}^{m} c_{is}(y_{i2} - \pi_{i2}) \\
&= 0, \quad (2\text{–}9)
\end{aligned}$$

for $s = 1, \cdots, p$. It is also easy to see that $E(Y_{i1} + Y_{i2}) = E(Y_{i+}) = \pi_{i1} + \pi_{i2}$. Thus, we have the likelihood equations

$$\sum_{i=1}^{m} c_{is} y_{i2} = \sum_{i=1}^{m} c_{is} \pi_{i2}, \quad (2\text{–}10)$$

$$y_{i+} = \sum_{j=1}^{2} y_{ij} = \sum_{j=1}^{2} \pi_{ij}, \quad (2\text{–}11)$$

where $s = 1, \cdots, p$.

Define the total count numbers as $n_{11} = \sum_{i=1}^{m} I\{Y_{i1} = 0, Y_{i2} = 0\}$, $n_{12} = \sum_{i=1}^{m} I\{Y_{i1} = 0, Y_{i2} = 1\}$, $n_{21} = \sum_{i=1}^{m} I\{Y_{i1} = 1, Y_{i2} = 0\}$, and $n_{22} = \sum_{i=1}^{m} I\{Y_{i1} = 1, Y_{i2} = 1\}$. Substitute $\frac{\exp(b_i^*)}{1+\exp(b_i^*)} + \frac{\exp(b_i^*+c_i^T\beta)}{1+\exp(b_i^*+c_i^T\beta)}$ in the right hand side of equation (2–11), then

(a) when $\hat{b}_i^* = -\infty$(i.e. $\hat{\pi}_{i1} = \hat{\pi}_{i2} = 0$), it refers the n_{11} subjects with $y_{i+} = 0$, i.e. $\{Y_{i1} = 0, Y_{i2} = 0\}$;

(b) when $\hat{b}_i^* = +\infty$(i.e. $\hat{\pi}_{i1} = \hat{\pi}_{i2} = 1$), it refers the n_{22} subjects with $y_{i+} = 2$, i.e. $\{Y_{i1} = 1, Y_{i2} = 1\}$;

(c) when $\hat{b}_i^* = -\dfrac{c_i^T\hat{\beta}}{2}$(i.e. $\hat{\pi}_{i1} + \hat{\pi}_{i2} = 1$), it refers the $(n_{12} + n_{21})$ subjects with $y_{i+} = 1$, i.e. $\{Y_{i1} = 0, Y_{i2} = 1\}$ or $\{Y_{i1} = 1, Y_{i2} = 0\}$.

By breaking $\sum_{i=1}^{m} P(Y_{ij} = 1)$ into components for the sets of subjects having $y_{i+} = 0$, $y_{i+} = 2$ and $y_{i+} = 1$, the right hand side of the likelihood equation (2–10) is

$$\begin{aligned}
&\sum_{i=1}^{m} c_{is}(I\{Y_{i1} = 0, Y_{i2} = 0\}\pi_{i2}(0) + I\{Y_{i1} = 1, Y_{i2} = 1\}\pi_{i2}(1) \\
&+ [I\{Y_{i1} = 1, Y_{i2} = 0\} + I\{Y_{i1} = 0, Y_{i2} = 1\}]\,\frac{\exp(\frac{c_i^T\hat{\beta}}{2})}{1+\exp(\frac{c_i^T\hat{\beta}}{2})}) \\
= \;& \sum_{i=1}^{m} c_{is}(I\{Y_{i1} = 1, Y_{i2} = 1\} + [I\{Y_{i1} = 1, Y_{i2} = 0\} + I\{Y_{i1} = 0, Y_{i2} = 1\}]\frac{\exp(\frac{c_i^T\hat{\beta}}{2})}{1+\exp(\frac{c_i^T\hat{\beta}}{2})}),
\end{aligned} \tag{2–12}$$

where $s = 1, 2, \cdots, p$. Meanwhile, the left hand side of (2–10) is

$$\begin{aligned}
&\sum_{i=1}^{m} c_{is}y_{i2} \\
=\;& \sum_{i=1}^{m} c_{is}I\{Y_{i2} = 1\} \\
=\;& \sum_{i=1}^{m} c_{is}(I\{Y_{i1} = 1, Y_{i2} = 1\} + I\{Y_{i1} = 0, Y_{i2} = 1\}).
\end{aligned} \tag{2–13}$$

Then we set that (2–12)=(2–13), that is

$$\begin{aligned} & \sum_{i=1}^{m} c_{is} I\{Y_{i1}=0,\, Y_{i2}=1\} \\ - & \sum_{i=1}^{m} c_{is}(I\{Y_{i1}=0,\, Y_{i2}=1\}+I\{Y_{i1}=1,\, Y_{i2}=0\})\frac{\exp(\frac{c_i^T\hat{\beta}}{2})}{1+\exp(\frac{c_i^T\hat{\beta}}{2})} \\ = & \; 0, \end{aligned} \tag{2–14}$$

for $s = 1, 2, \cdots, p$. Solving β in equation (2–14), we can show that $\hat{\beta}$ is the MLE.

Denote $\hat{\beta}_c(k)$ is the CMLE which is the solution to (2–6). If we set $k = 1$, then equation (2–6) becomes

$$\begin{aligned} & \sum_{i=1}^{m} c_{is} I\{Y_{i1}=0,\, Y_{i2}=0\} \\ - & \sum_{i=1}^{m} c_{is}(I\{Y_{i1}=0,\, Y_{i2}=1\}+I\{Y_{i1}=1,\, Y_{i2}=0\})\frac{\exp(c_i^T\beta)}{1+\exp(c_i^T\beta)} \\ = & \; 0, \end{aligned} \tag{2–15}$$

where $\hat{\beta}_c(1)$ is the solution (CMLE).

Comparing equations (2–14) and (2–15), it is easy to see that $\hat{\beta}_c(1) = \frac{\hat{\beta}}{2}$. Therefore, by Lemma 1, $\lim_{k\to\infty} \hat{\beta}_c(k) = \hat{\beta}$. Moreover, by Lemma 2, the convergence is monotonic. Since $\hat{\beta}_c(k)$ monotonically converge from $\frac{\hat{\beta}}{2}$ to $\hat{\beta}$ when k increase from 1 to ∞, then the range each element of $\hat{\beta}_c(k)$ is $\left[\min(\frac{\hat{\beta}_s}{2}, \hat{\beta}_s), \max(\frac{\hat{\beta}_s}{2}, \hat{\beta}_s)\right]$, for $s = 1, \cdots, p$ and $k = 1, 2, \cdots, \infty$.

For simple matched pair designs, we have the Corollary and Lemma 3 and Lemma 4 below as a special case for the Theorem.

Corollary. *For model (2–1) and the matched pairs design with $n_i = 2$, $X_{i1} = 0$ and $X_{i2} = 1$, let n_{12} be the number of observations with $Y_{i1} = 0$ and $Y_{i2} = 1$, and n_{21} be the number of observations with $Y_{i1} = 0$ and $Y_{i2} = 1$. Denote by $\hat{\beta}_c(k)$ the CMLE that maximizes (2–2) based on pretending that a pseudosample in which $w_i^* = 1$ and $w_{j|i}^* = k$*

is the observed sample. Then as $k \to \infty$, $\hat{\beta}_c(k)$ approaches $\hat{\beta}$ monotonically, where $\hat{\beta}$ is the standard MLE based on the original sample.

Lemma 3.

$$\lim_{k\to\infty} 2\log \frac{\sum_{r=0}^{k-1}\binom{k}{r}^2 \frac{k-r}{k} e^{(k-r)\beta}}{1+\sum_{r=0}^{k-1}\binom{k}{r}^2 \frac{r}{k} e^{(k-r)\beta}} = \beta. \tag{2–16}$$

Lemma 4. *When $\beta \geq 0$, $2\log \frac{\sum_{r=0}^{k-1}\binom{k}{r}^2 \frac{k-r}{k} e^{(k-r)\beta}}{1+\sum_{r=0}^{k-1}\binom{k}{r}^2 \frac{r}{k} e^{(k-r)\beta}}$ is non-increasing with $k \to \infty$; when $\beta < 0$, $2\log \frac{\sum_{r=0}^{k-1}\binom{k}{r}^2 \frac{k-r}{k} e^{(k-r)\beta}}{1+\sum_{r=0}^{k-1}\binom{k}{r}^2 \frac{r}{k} e^{(k-r)\beta}}$ is non-decreasing with $k \to \infty$.*

The proofs of Lemma 3 and Lemma 4 was first proved by He and Brumback [28]. Simplified proofs can be derived from the proofs of Lemma 1 and 2 by noticing that

$$\frac{\sum_{r=0}^{k}\binom{k}{r}^2 \frac{k-r}{k} e^{(k-r)\beta}}{\sum_{r=0}^{k}\binom{k}{r}^2 \frac{r}{k} e^{(k-r)\beta}} = \left\{1 - \frac{\sum_{r=0}^{k}\binom{k}{r}^2 (k-r) e^{(k-r)\beta}}{\sum_{r=0}^{k}\binom{k}{r}^2 k e^{(k-r)\beta}}\right\}^{-1} - 1.$$

Hence

$$\begin{aligned}\lim_{k\to\infty} 2\log \frac{\sum_{r=0}^{k}\binom{k}{r}^2 \frac{k-r}{k} e^{(k-r)\beta}}{\sum_{r=0}^{k}\binom{k}{r}^2 \frac{r}{k} e^{(k-r)\beta}} &= 2\log\left\{\left(1 - \frac{e^{\beta/2}}{1+e^{\beta/2}}\right)^{-1} - 1\right\} \\ &= \beta\end{aligned}$$

and the monotonicity of $\frac{\sum_{r=0}^{k}\binom{k}{r}^2 \frac{k-r}{k} e^{(k-r)\beta}}{\sum_{r=0}^{k}\binom{k}{r}^2 \frac{r}{k} e^{(k-r)\beta}}$ is the same as $\frac{\sum_{r=0}^{k}\binom{k}{r}^2 (k-r) e^{(k-r)\beta}}{\sum_{r=0}^{k}\binom{k}{r}^2 k e^{(k-r)\beta}}$.

Denote $f_k(\beta) = 2\log \frac{\sum_{r=0}^{k-1}\binom{k}{r}^2 \frac{k-r}{k} e^{(k-r)\beta}}{1+\sum_{r=0}^{k-1}\binom{k}{r}^2 \frac{r}{k} e^{(k-r)\beta}}$. We used Maple (version 12) to construct Figure 2-1 and Figure 2-2 to illustrate that $\lim_{k\to\infty} f_k(\beta) = \beta$ and that the convergence is monotone. One observes that for $k = 1$, $f_1(0.5) = 1$ and $f_1(-0.5) = -1$. This coincides with equation (2–16), in which $f_1(\beta) = 2\beta$; in other words, $\hat{\beta}_c(1) = \log \frac{n_{21}}{n_{12}}$, as we would expect of the conditional maximum likelihood estimator assuming no replication ($k = 1$).

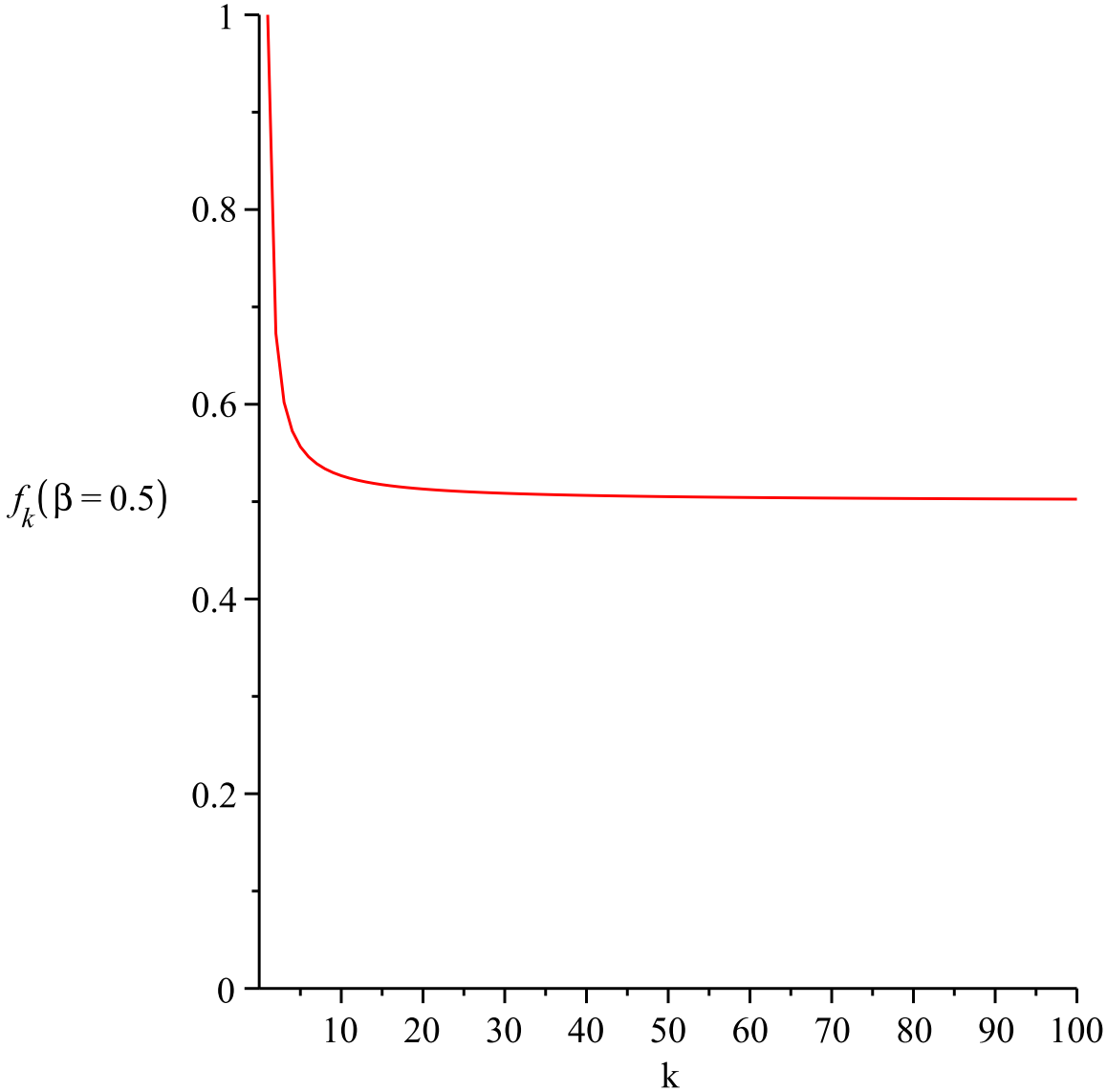

Figure 2-1. $f_k(\beta)$ as a function of k for $\beta = 0.5$

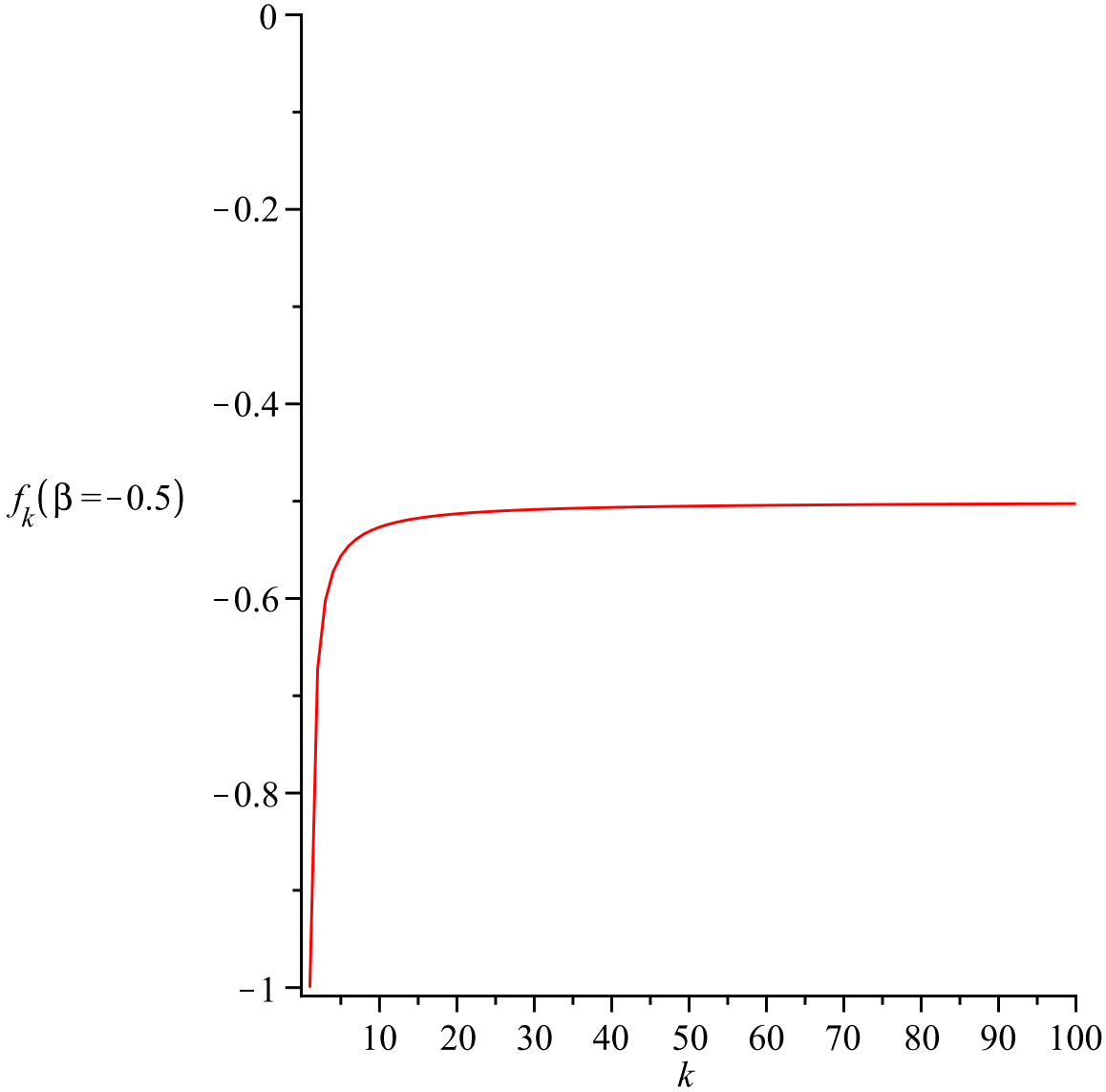

Figure 2-2. $f_k(\beta)$ as a function of k for $\beta = -0.5$

2.4 Simulation Study

We conducted a simulation study using a matched pairs design with two covariates. We sampled $m = 1000$ clusters and we let $n_i = 2$. We let the first covariate X_{ij} be Bernoulli with probability $p_{xij} = \text{logit}^{-1}(u_i)$, where u_i, $i = 1, \ldots, m$ is i.i.d. $N(0, 1)$. We let the second covariate Z_{ij} be i.i.d. standard normal distribution $N(0, 1)$. Let δ_i, $i = 1, \ldots, m$ also be i.i.d $N(0, 1)$ (independent of the u_i), $\bar{X}_i = (1/n_i)\sum_j X_{ij}$ and $\bar{Z}_i = (1/n_i)\sum_j Z_{ij}$,

and we simulated $b_i = -5\bar{X}_i + 3\bar{Z}_i + \delta_i$. Finally we simulated binary responses $Y_{ij} = \text{logit}^{-1}(0.5X_{ij} + 0.8Z_{ij} + b_i)$. The true $\boldsymbol{\beta}$ is $\boldsymbol{\beta} = (0.5, 0.8)$. We repeated the simulation 100 times.

Again denote by $\hat{\boldsymbol{\beta}}_c(k)$ the conditional maximum likelihood estimator that maximizes (2–2) based on pretending that a pseudosample in which $w_i^* = 1$ and $w_{j|i}^* = k$ is the observed sample. Let $\hat{\boldsymbol{\beta}}$ denote the unconditional MLE based on the original sample. Table 1 presents the average and standard deviation of $\hat{\boldsymbol{\beta}}_c(k)$ with respect to the 100 simulations. For comparison, the average (standard deviation) of $\hat{\boldsymbol{\beta}}$ is (1.0814, 1.6508)(standard deviations are 0.3838 and 0.1935 respectively). One readily observes that $\hat{\boldsymbol{\beta}}_c(k)$ tends towards $\hat{\boldsymbol{\beta}}$ monotonically as k increases.

Table 2-1. Results of simulation study with $\boldsymbol{\beta} = (0.5, 0.8)$. The MLE is $\hat{\boldsymbol{\beta}} = (1.0814, 1.6508)$ with $SD(\hat{\boldsymbol{\beta}}) = (0.3838, 0.1935)$.

k	$\hat{\beta}_c$	$SD(\hat{\beta}_c)$
1	(0.5407,0.8254)	(0.1919,0.0967)
2	(0.7243,1.1044)	(0.2550,0.1201)
3	(0.8148,1.2422)	(0.2866,0.1334)
4	(0.8691,1.3252)	(0.3058,0.1424)
5	(0.9055,1.3808)	(0.3187,0.1490)
6	(0.9316,1.4205)	(0.3281,0.1541)
7	(0.9511,1.4503)	(0.3351,0.1581)
8	(0.9662,1.4735)	(0.3406,0.1615)
9	(0.9783,1.4921)	(0.3450,0.1642)
10	(0.9882,1.5072)	(0.3487,0.1666)
15	(1.0187,1.5540)	(0.3599,0.1744)
20	(1.0343,1.5782)	(0.3658,0.1789)
50	(1.0628,1.6221)	(0.3767,0.1876)
100	(1.0722,1.6366)	(0.3802,0.1906)

2.5 Discussion

We have shown that for the matched-pairs design and a logistic regression model with an intercept for each pair, replicating observations k times and maximizing the conditional likelihood results in an estimator whose limit as k approaches infinity is exactly identical to the MLE with a fixed effect for each pair. We presented results of a simulation study and figures to illustrate our results. One implication of the result is that the use of conditional logistic regression with a pseudosample for complex survey designs will sometimes lead to a biased estimator.

It would be interesting to further investigate whether the equivalence of infinitely replicated conditional logistic regression with unconditional logistic regression extends to general designs. The approach taken in this main result section extend to general design for matched pairs with binary outcomes. An alternative approach, as mentioned in the introduction section, would be to approximate the conditional maximum likelihood estimator with a generalized linear mixed model estimator based on a smooth mixing distribution having positive support on the entire real line. Then, large-scale replication would lead to a GLMM estimator tending towards treating the random effects as fixed

effects (because the replicated observations would dominate the Bayesian prior), and treating the random effects as fixed effects yields the unconditional MLE. Previous authors have identified some results on the exact equivalence of CMLEs with GLMM estimators in the special case of the Rasch models [52]. Lindsay et al. [37] studied estimation technologies in Rasch models using conditional likelihood methods, while Rice ([54]and [55]) analysed the equivalence between conditional likelihood method and random effect models for Rasch models and matched-pair case-control studies. For these results to be useful even for constructing a proof alternative to the one presented here for the matched-pair design, we would need to know not only the existence of a mixing distribution yielding exact equivalence, but also about certain details of that mixing distribution (such as its smoothness and support) for $k : k$ matched studies. However, for $k : k$ matched studies, the form of mixing distribution remains unknown [55]. Yet another possibility would be attempt using the result in Proposition 3.2 of Severini [57] to approximate the conditional likelihood function to the third order with a generalized linear mixed model based on a complicated mixing distribution. This would lead to an asymptotic equivalence rather than an exact equivalence.

CHAPTER 3
ADJUSTING FOR CONFOUNDING BY NEIGHBORHOOD USING COMPLEX SURVEY DATA

3.1 Outline

Recently, Graubard and Korn [20] have adapted an older method to complex survey data. The method is based on a weighted pseudo-likelihood, in which the contribution from each neighborhood involves all pairs of cases and controls in the neighborhood. In this chapter, we implement their method by translating the pairwise pseudo-conditional likelihood into an equivalent ordinary weighted log-likelihood formulation. We also show that it corresponds to an equivalent ordinary weighted log-likelihood formulation with binary outcomes. We explain how to program the method using standard software for ordinary logistic regression with complex survey data. We then apply the method to 2009 National Health Interview Survey (NHIS) public use data, to estimate the effect of education on health insurance coverage, adjusting for confounding by neighborhood.

3.2 Introduction

In Chapter 1, we considered the problem of adjusting for confounding by neighborhood of an individual exposure effect on a binary outcome, using complex survey data in which neighborhoods are sampled with unequal probabilities, as are individuals within neighborhoods [14]. It is motivated by an analysis of NHIS public use data to estimate the effect of education on mammography screening, adjusting for confounding by neighborhood. When only some neighborhoods are sampled, the problem of adjusting for confounding by neighborhood is one of adjusting for confounding by cluster. In Chapter 1, we have examined three categories of methods: ordinary logistic regression for survey data, with either no effect or a fixed effect for each cluster; conditional logistic regression extended for survey data; and generalized linear mixed model (GLMM) regression for survey data. Because the conditional log-likelihood cannot be weighted to produce a pseudolikelihood, conditional maximum likelihood was applied to a pseudosample constructed by replicating observations according to integer-scaled

versions of the weights. The GLMM method utilized the generalized linear latent and mixed model (GLLAMM) Stata software of Rabe-Hesketh and Skrondal [51] combined with an adaptation for survey data of the 'poor man's' aproposed by Neuhaus and Kalbfleisch [44] and discussed by Neuhaus and McCulloch [46] and Brumback et al. [14]. The approach models the association of the latent neighborhood effect with the neighborhood-level vector of individual-level exposures using a Gaussian model linear in the neighborhood mean of the individual exposures. The GLLAMM software implements maximum pseudolikelihood with complex survey data. In Chapter 1, we showed that these methods worked in some well defined settings. But they failed dramatically when the sampling bias was strong and the sample sizes of the neighborhoods were small. This interesting finding motivates us to find a new method that works properly when the sampling bias is strong.

Subsequently, Graubard and Korn [20] showed how to adapt to complex survey data a method due to Liang [35] for adjusting for confounding by cluster. Their adaptation achieves a consistent estimator even when neighborhood sample sizes are small and the selection bias is strongly informative. The method is based on weighted pseudolikelihoods, where the contribution from each neighborhood involves all pairs of cases and controls in the neighborhood. The case-control pairs are treated as if they were independent, a pairwise pseudo-conditional likelihood is thus derived, and then the corresponding score equation is weighted with inverse-probabilities of sampling each case-control pair. In this chapter, we simplify the implementation of Graubard and Korn's [20] approach, using a method due to Breslow et al. [12] which translates the pairwise pseudo-conditional likelihood into an equivalent ordinary weighted log-likelihood; Agresti [1] (Section 10.2.6 and Section 4) described the approach of Breslow et al. [12]. Using the simplified implementation, we show how to easily program the method using software for ordinary logistic regression for complex survey data (e.g. SAS PROC SURVEYLOGISTIC.) We also broaden the sampling scenarios considered by Graubard

and Korn [20]. We show that the method applies even to sampling situations in which the primary sampling units (PSUs) are nested within neighborhood, as is the case with survey designs such as the Behavioral Risk Factor Surveillance System (BRFSS) survey [53]. We demonstrate the validity of our simplified implementation by applying it to the simulation considered by Brumback et al. [14] for which the previous methods failed; the new method performs beautifully. We also apply the new method to an analysis of 2009 NHIS public-use data [10], to estimate the effect of education on health insurance coverage, adjusting for confounding by neighborhood. We operationalize neighborhood as a person's household, which is present in the NHIS public-use files.

In next several sections, we present our simplified implementation method and the broaden the sampling scenarios where the method works. In Section 3.3 we describe the motivating NHIS example; Section 3.4 outlines the modeling framework; and Section 3.5 details the approach to estimation with complex survey data and the simplified implementation. We present simulation results in Section 3.6, and in Section 3.7 we analyze the NHIS data as an application. Section 3.8 concludes with a discussion.

3.3 Motivating NHIS Example

Our research has been motivated by an analysis of 2005 NHIS data discussed by Brumback et al. [14]. It is a study of individual-level and small neighborhood-level predictors of health-related outcomes such as mammography screening and health insurance status [10]. Individual-level predictors that we have studied include education, age category, and race/ethnicity. Neighborhood-level predictors include measured census-tract variables such as poverty and segregation indices, as well as the high-dimensional categorical variable that is neighborhood itself. In the course of our research, the question came up to our mind how could we adjust for confounding of individual level variables by unmeasured neighborhood factors using the high-dimensional categorical variable? In this section, for expository purposes, we will focus our attention on the effect of dichotomized education (more than high school versus high school or

less) on health insurance status (covered or not covered) adjusting for confounding by neighborhood, operationalized as household, because census-tract or the secondary sampling unit (SSU) information is not included in the public-use file.

We used the 2009 NHIS data and considered only adults over the age of 18; the total sample size was 64,616 adults. Then we excluded individuals with missing or unknown responses to the education and insurance coverage questions; this resulted in a final sample size of 62,904 adults nested in 33,429 households ranging in size from 1 to 9. Ninety-five percent of the households in the final sample had three or fewer included adult respondents. We used the interim annual survey weights (WTIA in the public-use file), which do not include the final poststratification adjustment. Poststratification adjusts weights associated with the sampled data so that the joint distribution of a set of post-stratifying variables matches the known population joint distribution ([31], [39]). Poststratification adjustments may interfere with the computation of the pairwise weights that we need for the methodology, and we will discuss this in detail in Section 3.5.

3.4 Modeling Framework

Let $i = 1, \cdots, M$ index all neighborhoods in the population, and $j = 1, \cdots, N_i$ index all individuals in the population within a given neighborhood. Let Y_{ij} be a binary variable indicating health insurance coverage for person j in neighborhood i, and correspondingly let X_{ij} similarly indicate education level. Then a simple model for the effect of exposure X_{ij} on outcome Y_{ij} that accounts for confounding by a general household effect b_i is

$$h(E(Y_{ij}|X_i, b_i)) = X_{ij}\beta + b_i, \tag{3–1}$$

where h is the logit link function and $X_i = (X_{i1}, \cdots, X_{iN_i})$ indicates all the exposure information within neighborhood i.

A causal interpretation of β in model (3–1) would be based on a potential outcome framework with Y_{ij}^x being the potential outcome to setting $X_{ij} = x$. If household was the

only confounder, we would assume

$$\{Y_{ij}^{x}\} \amalg X_{ij}|b_i, \tag{3–2}$$

where $\{Y_{ij}^{x}\}$ denotes the set of potential outcomes defined by all possible values of x. Then we would also need a consistency assumption, which states that the potential outcomes Y_{ij}^{x} are well-defined and that those for a given individual do not depend on the realized exposures for other individuals. With these assumptions, β of model (3–1) is identical to β of the following causal model

$$h(E(Y_{ij}^{x}|b_i)) = x\beta + b_i. \tag{3–3}$$

The causal model (3–3) is set up based on the potential outcomes $\{Y_{ij}^{x}\}$ regarding all possible values of x. In Chapter 1, we discussed, in some conditions, that the consistency assumption and causal model (3–3) could be dubious. Therefore we will focus our interest in model (3–1) rather than the causal model (3–3). Model (3–1) retains interest even when a causal interpretation is untenable; one may wish to know how much of the association between X_{ij} and Y_{ij} can be accounted for by unmeasured household-level factors.

3.5 Estimation with Complex Survey Data

Denote the probability of selecting individual j from household i into the sample as p_{ij} and the corresponding inverse-probability weight as w_{ij}.

If we had an ordinary cluster sample according to the models in Section 3.4, we could use conditional logistic regression to estimate β. Conditional logistic regression conditions on sufficient statistics for the b_i, and finds the maximum likelihood estimator based on the conditional likelihood [1]. Estimation and inference are implemented in, for example, SAS PROC LOGISTIC using the stratum statement. As demonstrated by Graubard and Korn [20], this method does not readily generalize to the complex survey sample setting. Following their exposition, order the observations so that, within cluster

i , the first K_i individuals have $Y_{ij} = 1$ and the remaining $N_i - K_i$ have $Y_{ij} = 0$. Then the conditional likelihood contribution from cluster i is

$$L_i = \frac{\prod_{j=1}^{K_i} \exp(X_{ij}\beta)}{\sum_c \prod_{j=1}^{K_i} \exp(X_{ic(j)}\beta)}, \tag{3–4}$$

where the sum in the denominator is over all N_i choose K_i ways (indexed by c) of selecting K_i elements from the set $\{1, 2, \cdots, N_i\}$ [1]. $c(j)$ represents the jth of the K_i elements selected by c. As given by Graubard and Korn [20], the contribution to the score equation from cluster i thus has the form

$$S_i(\beta) = \sum_{j=1}^{K_i} X_{ij} - \frac{\sum_c (\sum_{j=1}^{K_i} X_{ic(j)})(\prod_{j=1}^{K_i} \exp(X_{ic(j)}\beta))}{\sum_c \prod_{j=1}^{K_i} \exp(X_{ic(j)}\beta)}. \tag{3–5}$$

Then we have the total score equation, $S(\beta) = \sum_i^M S_i(\beta)$. One can observe that it is not a simple sum of contributions from each person in the population indexed by i and j. Therefore, it is not possible to weight contributions of such a sum with weights w_{ij}, as would be done with pseudolikelihood estimation.

Following Liang [35], Graubard and Korn [20] resolved the problem by considering the cluster-specific pairwise conditional likelihood

$$L_{L_i} = \prod_{j=1}^{K_i} \prod_{l=Ki+1}^{N_i} \frac{\exp(X_{ij}\beta)}{(\exp(X_{ij}\beta) + \exp(X_{il}\beta))}. \tag{3–6}$$

This conditional likelihood (3–6) would arise in cluster i if all possible pairs (j, l) of observations with $Y_{ij} = 1$ and $Y_{il} = 0$ were treated as if they were independent (see Agresti [1], Section 10.2.6). Graubard and Korn [20] give the score equations corresponding to L_{L_i} as

$$S_{L_i} = \sum_{j=1}^{K_i} \sum_{l=K_i+1}^{N_i} \left(X_{ij} - \frac{X_{ij} \exp(X_{ij}\beta) + X_{il} \exp(X_{il}\beta)}{\exp(X_{ij}\beta) + \exp(X_{il}\beta)} \right). \tag{3–7}$$

The score Equation (3–7) sums over all possible pairs (j, l) of positive and negative outcomes for cluster i. The total score equation is then $S_L(\beta) = \sum_{i=1}^M S_{L_i}(\beta)$.

Next we estimate (3–7) using complex survey data. We denote the probability of observing pair (j, l) as q_{ijl}, and we let w^*_{ijl} denote its reciprocal. For the NHIS data, the probability of observing a pair from the same cluster is the same as the probability that we observe one member of that cluster, because all adults in a household were sampled. In other words, $w^*_{ijl} = w_{ij} = w_{il}$. If we were using BRFSS data, in which individuals are sampled roughly independently of one another, the probability of observing a pair would be approximately the product of the probabilities that we observe each member of the pair. For our simulation in Section 3.6, the probability of the pair is also the product of the probabilities, $w^*_{ijl} = w_{ij} \times w_{il}$. We assume that there are m sampled clusters, k_i sampled individuals with $Y_{ij} = 1$ in cluster i, and $(n_i - k_i)$ sampled individuals with $Y_{ij} = 0$ in cluster i.

Using the pairwise weights w^*_{ijl}, we estimate S_{L_i} by

$$\hat{S}_{L_i} = \sum_{j=1}^{k_i} \sum_{l=k_i+1}^{n_i} w^*_{ijl} \left(X_{ij} - \frac{X_{ij} \exp(X_{ij}\beta) + X_{il} \exp(X_{il}\beta)}{\exp(X_{ij}\beta) + \exp(X_{il}\beta)} \right), \tag{3–8}$$

and correspondingly we estimate S_{L_i} by $\hat{S}_L = \sum_{i=1}^{m} \hat{S}_{L_i}$.

3.5.1 Simplified Implementation

Graubard and Korn [20] used special programming to maximize $\hat{S}_L$ and compute standard errors and confidence intervals. However, if one makes use of the correspondence between conditional maximum likelihood for matched pairs and ordinary maximum likelihood first proposed by Breslow et al. [12], the programming greatly simplifies. Consider the contribution to L_{L_i} from pair (j, l), in which $Y_{ij} = 1$ and $Y_{il} = 0$, which is given by

$$P(Y_{ij} = 1, Y_{il} = 0 | Y_{ij} + Y_{il} = 1) = \frac{\exp(x_{ij}\beta)}{\exp(x_{ij}\beta) + \exp(x_{il}\beta)}. \tag{3–9}$$

Agresti ([1], Section 10.2.6) explains that if we divide the numerator and denominator by $\exp(x_{ij}\beta)$, the equation has the form of logistic regression with no intercept, with $Y^*_i(j, l) = 0$ as a constant outcome, and predictor values $X^*_i(j, l) = X_{il} - X_{ij}$. Then

Equation (3–9) can be revised as a ordinary logistic regression form

$$P(Y_{ij} = 1, Y_{il} = 0 | Y_{ij} + Y_{il} = 1) = P(Y_i^*(j, l) = 0 | Y_{ij} + Y_{il} = 1) = \frac{1}{1 + \exp(x_i^*(j, l)\beta)}. \tag{3–10}$$

This gives rise to an estimating equation S_L^* that has the form of weighted ordinary logistic regression.

Therefore, SAS PROC LOGISTIC (without the stratum statement) could be used with $Y_i^*(j, l)$ and $X_i^*(j, l)$ to estimate β when the weights w_{ijl}^* are all equal to one, but the estimated sampling variability would be inconsistent because the pairs (j, l) are not independent. A convenient feature of software for survey data, such as SAS PROC SURVEYLOGISTIC, is that consistent estimators of sampling variability are obtained provided that the specified PSUs are independent. Furthermore, the weights w_{ijl}^* need not equal one, and the procedures can accommodate complex multistage sampling. To accommodate complex multistage sampling designs, we can use a sandwich estimator for the variance of the estimating equation S_L^* provided that the PSUs are independent [58]. We note that one can partition the data set into alternative PSUs and still obtain consistent estimators of the variance, provided that two alternative PSUs do not contain observations belonging to a single PSU, and also that the alternative PSUs reside in the same stratum as the single PSU nested within them. Therefore, we can use the simplified procedure even when the PSU represents an individual, as the same with the BRFSS survey data. In other words, for the BRFSS data, we can let the alternative PSU be the neighborhood, provided that the neighborhood lies within a single stratum. Even if neighborhood overlaps strata, we can obtain conservative estimates of variability by ignoring the stratification because incorporating stratification can only reduce the estimated variance. For the NHIS data that we analyze in this chapter, neighborhood (i.e. household) is nested within PSU, and therefore we can use the public-use PSUs in conjunction with the CLUSTER statement of PROC SURVEYLOGISTIC.

Unlike SAS PROC LOGISTIC, SAS PROC SURVEYLOGISTIC will not work when all of the outcomes are equal to zero. However, following the above argument in Agresti [1], we can handle this problem by setting any fraction (e.g. those corresponding to the first thousand households) of the paired observations to have $Y_i^*(j,l) = 1$ and $X_i^*(j,l) = X_{ij} - X_{il}$. In the appendix, we will provide simple SAS code for the data management and the use of PROC SURVEYLOGISTIC in SAS software.

All inference will be directed at superpopulation [58] parameters. Suppose the population of M neighborhoods and N_i individuals per neighborhood in model (3–1) is conceptualized as a random two-stage sample from the superpopulation, for which a version of model (3–1) holds but for infinite M and N_i. In the first stage, M neighborhoods are selected completely at random from a hypothetically infinite population of neighborhoods, and in the second stage, N_i individuals are selected completely at random from a hypothetically infinite population of individuals within neighborhood.

3.5.2 Extension to Multiple Individual-Level Covariates

The above methods extend very simply to handle multiple individual-level covariates in model (3–1). For p dimensional covariates, model (3–1) is then revised as

$$h(E(Y_{ij}|\mathbf{X}_i, b_i) = \mathbf{X}_{\mathrm{ij}}{}^T\boldsymbol{\beta} + b_i, \tag{3–11}$$

where $\mathbf{X}_{\mathrm{ij}} \in \mathbb{R}^p$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\mathbf{X}_i$ denotes all the covariates information within cluster i, i.e. $\mathbf{X}_i = (\mathbf{X}_{\mathrm{i1}}, \cdots, \mathbf{X}_{\mathrm{iN_i}})^T$.

We form a covariate, $\mathbf{X}_i^*(j,l) = \mathbf{X}_{il} - \mathbf{X}_{il}$, analogous to $X_i^*(j,l)$ for each of the multiple covariates. Then the implementation and interpretation can be proceeded as described in Section 3.5.1.

3.5.3 Specifying the Pairwise Weights in Practice

In practice, specifying the pairwise weights may be challenging or require additional assumptions for validity. With our NHIS example, we are using the survey design

weights, which neither include the final poststratification adjustment for unit nonresponse nor adjustments for item nonresponse. This leads to accurate inference provided that the nonresponse is not informative.

An alternative possibility would be to multiply w^*_{ijl} by the factor a_{ijl} , which is intended to represent the inverse-probability of complete data from pair (ij, il) conditional upon sampling pair (ij, il) and on the survey design variables. In practice, a_{ijl} would likely be modeled as $a_{ij}a_{jl}$, where a_{ij} is either a poststratification adjustment for individual ij or the product of the poststratification adjustment and a weight for item nonresponse. This method would lead to accurate inference provided that the a_{ij} accurately represents the inverse-probability of complete data from person ij conditional on having sampled person ij and on the design variables, and that the conditional probability of complete data from person ij is independent of that for person il. Typically, however, a_{ij} is an ad hoc adjustment that does not truly represent the inverse-probability of complete data from person ij conditional on having sampled person ij and on the design variables. For example, the nonresponse probabilities might actually depend on the survey design variables, instead of being constant across them as is a typical poststratification adjustment factor. However, for situations in which nonresponse is believed to be informative, it may be preferable to use the factor $a_{ij}a_{il}$ and invoke the assumption that it approximates the true a_{ijl}.

3.6 Simulation Study

In this section, we demonstrated the validity of the new method and implementation using the strong sampling bias simulation settings of Chapter 1 (Brumback et al. [14]), where previous methods failed.

First, we simulated the population with M arbitrarily large (in practice, we needed only set M=m=1000) and $N_i = 1000$ for each i. We let X_{ij} be independent (given u_i) Bernoulli random variables with probability expit(u_i), where u_i are i.i.d. $N(0, 1)$. We let $b_i = -5\bar{X}_i + \delta_i$,with δ_i i.i.d. $N(0, 1)$. Finally, we generated Y_{ij} according to model (3–1)

with our target of estimation β set equal to 0.5. Second, we sampled from the population with strongly biased sampling scheme. We first divided the individual observations into concordant ($C = 1$) and discordant ($C = 0$) groups based on whether $Y_{ij} = X_{ij}$ or not. We finally included observations into the sample with independent probability 0.002 if $C = 1$ and 0.004 if $C = 0$. Then we included $m = 1000$ neighborhoods into the sample, and n_i varied for each cluster due to our sampling scheme. Observations were given survey weights $w_{ij} = 2$ if $C = 1$, and $w_{ij} = 1$ if $C = 0$. Owing to independent sampling within neighborhoods, $w^*_{ijl} = w_{ij} w_{il}$. The simulation was repeated 100 times to study the performance of the method in estimating β (true value=0.5).

3.6.1 Results

The mean of the estimates of β was 0.5108 with a standard deviation of 0.18 and a range of (0.16, 0.99). This indicates that the method and the implementation perform very well indeed, especially when compared with the methods considered in Chaper 1 (Brumback et al. [14]). In Chapter 1, for the same simulation settings, for conditional logistic regression with a pseudosample the mean of the estimates was 0.10 (SD=0.15, range=(-0.22,0.47)), and for the GLLAMM method, the mean of the estimates was -0.27 (SD=0.18, range=(-0.62,0.16)) (see Table (1-2)). Comparing these simulation results, our implemented new method is obviously superior to the other two methods under the strongly biased sampling scheme.

3.7 Application to NHIS Data

Next we applied the new method and implementation to the NHIS data to estimate the effect of education on health insurance coverage, adjusting for confounding by household. The adjusted odds ratio was 1.23 with a 95 percent confidence interval of (1.10, 1.38). For comparison, we also computed the crude odds ratio using ordinary logistic regression for complex survey data [32]; it was 1.27 with a 95 percent confidence interval of (1.18, 1.37). Thus, at the first glance, there appears to be not much confounding by household, because the adjusted odds ratio estimate is similar to

that of the crude odds ratio. As we might expect, adjusting for household attenuates the estimated association between education and health insurance coverage, although the attenuation is slight. Unlike the risk difference and relative risk, the odds ratio is not collapsible even assuming no confounding [25], and our outcome is relatively prevalent (the estimated percent of adults with health insurance coverage is 82.9 percent with a 95 percent confidence interval of (82.3, 83.4), so that the odds ratio does not approximate the relative risk. Conversely, that the household-specific odds ratio and the population-averaged odds ratio are equal does not imply that there is no confounding. In Chapter 1 (Brumback et al. [14]), we discussed a method for estimating a population-averaged odds ratio that is adjusted (i.e. standardized) for confounding by cluster; however, that method utilized a GLMM approach rather than the conditional logistic regression approach considered in Chapter 1. Estimating a population-averaged odds ratio that is standardized for confounding by cluster using complex survey data is still an open problem.

3.8 Discussion

We have considered the problem of adjusting for confounding by cluster in the context of complex multistage sampling and a binary outcome. We have achieved our goal of finding a method that works in the setting of the strongly biased sampling simulation first considered by Brumback et al. [14]. We have also succeeded in implementing the method using standard software and basic data management. Furthermore, we have pointed out that the applicability of the method is broader than that considered by Graubard and Korn [20]. Nevertheless, it would be of further interest to find a method that would allow us to estimate the distribution of the random cluster effects conditional on cluster-level aggregate functions of the individual covariates. In other words, whereas we have found a generalization of conditional logistic regression for complex survey data, we have not found a generalization of GLMMs for complex

survey data that works when the sampling bias is strong and the cluster sample sizes are small. Further research is still needed.

Additionally, the method used in this chapter requires a knowledge of the probability that a given pair of observations within a neighborhood is selected into the sample. This probability was easily obtained for our operationalization of neighborhood, because every member of a household is sampled for the NHIS. However, in Chapter 1 (Brumback et al. [14]), we operationalized neighborhood as SSU, and not every person in an SSU is sampled. In their application, individuals in the data set were from different households. Based on the superpopulation model that supposes that the superpopulation neighborhood is infinite in size, one might plausibly assume that individuals j and l were sampled independently from the superpopulation, given that their SSU was sampled. For the NHIS application in the this chapter, if we had to define neighborhood as SSU, we would need to know whether individuals j and l were in the same household or not, which would not be a problem, because this information is available to us. For other applications, however, it could be the case that the defined neighborhood is bigger than the clusters entailed by the final clustering stage of the survey design, and that those clusters are furthermore not identified in the data set (e.g. as with many public-use data sets). In these cases, the methodology we described would not be possible to implement.

CHAPTER 4
ESTIMATING THE EFFECT OF CLUSTER-LEVEL ADHERENCE ON AN INDIVIDUAL BINARY OUTCOME WITH A COMPLEX SAMPLING DESIGN

4.1 Outline

In this chapter, we wish to estimate the effect of school-level adherence on individual absenteeism in the context of a school-based water, sanitation, and hygiene intervention in Western Kenya. Schools within strata were disproportionately sampled and randomized to one of three interventions. Next, students within schools were disproportionately sampled and measured for outcomes such as absenteeism. We use double inverse-probability weighting to adjust for the disproportionate sampling and the association of individual-level confounders with randomization. We develop and apply methods based on structural nested models to estimate effects of adherence assessed in terms of relative risks, using school-level randomization as an instrumental variable and using the double weights to adjust for complex sampling and individual-level confounding.

4.2 Introduction

In this chapter, we want to estimate the effect of intervention arms the school received on the pupil absenteeism in a cluster-randomized trial of school-based water, sanitation and hygiene (SWASH) project conducted in Western Kenya, from 2007-2008. The study was conducted in 86 schools within two areas in Western Kenya, Rachuonyo District and Suba District. The 86 schools within two areas were disproportionately sampled and randomized to one of three interventions; the intervention arms are

1. HP & WT: the schools will be provided with hygiene promotions and water treatment; here the hygiene promotions is considered as providing soap.
2. HP & WT + sanitation: the schools will not only be provided with hygiene promotions and water treatment, but also sanitation intervention, such as latrine construction.
3. Control: the schools will be provided no interventions above during the study period. They will receive all the intervention package at the end of the study.

Students within schools were disproportionately sampled into the study. Then the staff of the study recorded whether students were absent from school on the day of measurement along with some individual-level and household-level characteristics. Thus, the final dataset consists of 1029 pupils from 86 schools within Rachuonyo District and Suba District in Western Kenya. In this chapter, we would like to apply the instrumental variable method with a structural nested model on SWASH project.

4.2.1 Instrumental Variable

The method of instrumental variable (IV) is widely used in statistics, econometrics, epidemiology, genetics and related disciplines. The conception of IVs was first invented by Wright [66] in 1928 (See [59] and [3]). Wright investigated the elasticities of supply and demand for product, like flaxseed (the source of linseed oil). In the book, he explained how to use ”curve shifters” (now called IV) to address the existence of confounding: “Such additional factors may be factors which (A) affect demand conditions without affecting cost conditions or which (B) affect cost conditions without affecting demand conditions.” IV method could be used to adapt to the observational study to adjust for unmeasured confounding. Comparatively, most methods are used to control for measured confounders, which are known and included in the model as part of the covariates. After its first invention, IV method was mainly used in macroeconomics and microeconomics.

In recent decades, IV methods began to enter the area of health science to estimate the causal effect of treatment or gene on the outcome adjusting for unmeasured confounding. Angrist et al. [4] outlined the framework of causal inference using IV in the setting of randomized treatment design when compliance is not ignorable. Greenland [21] gave the epidemiologists a brief introduction of IV method and assumption with a causal diagram for treatment randomization experiment. Hernán and Robins [30] discussed IV method applied to estimate parameters of several causal models, including structural equation models and structural mean models. In genetics field, IV method is

usually called ”Mendelian randomization” [65] and has been grown in recent 10 years. Recently, Glymour et al. [18] discussed approaches to evaluate assumptions for valid IV associated with a genome wide association study. In application of IV method, many researchers conduct two-stage least square (2SLS) with the instrument, endogenous variable and the outcome (See [64], [60], [50], [13], [5] and [6]). Just as the name implies, one conducts 2SLS IV by applying standard ordinary least squares (OLS) in two stages. In the first stage, one conducts OLS regression of the endogenous variable on the instrument and estimate the predicted endogenous variable for each subject. In the second stage, one applies OLS regression of the outcome on the predicted endogenous variable. The estimated parameter in the second stage is the IV estimator. However, the 2SLS IV doesn't always work. The 2SLS IV can only be applied on some linear causal models. For example,

$$\begin{aligned} Y &= X\beta + U_1 \\ X &= Z\theta + U_2 \end{aligned} \tag{4–1}$$

where Y is the outcome, Z is the instrument variable, X is the endogenous variable. Z is uncorrelated with the error U_1 and U_2(exclusive assumption);and Z is correlated with X. Bun and Windmeijer [16] compared the bias of 2SLS estimator under several scenarios with different number and strength of instruments. The 2SLS estimator can perform poorly with weak instruments or many instruments or both on finite samples.

The advantage of IV is obvious in that it can be used to adjust for unmeasured confounding for observational studies. Several articles compared IV methods with other commonly used methods in observational studies, including intent to treat analysis, as-treat analysis, per-protocal analysis ([38] and [61]). Many researchers realize that the IV method relies strongly on the assumptions and discuss the limitation of IV (See [11] and [42]).

$$\begin{tikzcd} Z \arrow[r] & A \arrow[r] & Y \\ & U \arrow[u] \arrow[ur] & \end{tikzcd}$$

Figure 4-1. DAG represents the causal relationship between variables.

Next, we will describe the IV assumptions and a particular causal directed acyclic graph (DAG) mostly represented with IV method. Define U is the unmeasured confounder, Z is the randomization of the treatment (instrument), A is the endogenous variable (such as compliance) and Y is the outcome we observed. Then we can represent the relationship between the variables using a directed acyclic graph (DAG) as Figure 4-1 [21]. The assumptions of IV is stated in [24]

1. Z is independent of U;
2. Z is associated with A;
3. Z is independent of Y given U and A.

We will discuss the modified assumptions with regard to application on School-based Water, Sanitation, and Hygiene (SWASH) project in the later section.

4.2.2 Structural Nested Models

Structural nested models were first developed by Robins [56], where it can be adapted with many link functions, such as linear link, logit link. Vansteelandt and Goetghebeur [62] extended and applied structural nested models to some instrumental-variables settings, but not including complex sampling designs.

Below are the notations we will use in the future.

i: cluster in the data, $i = 1, \cdots, m$;

j: individual j within the school i, $j = 1, \cdots, n_i$;

Y_{ij}: binary individual-level outcome;

Z_i: cluster level randomization;

A_i: cluster level adherence; its degree of freedom is no greater than the degree of freedom of Z_i; otherwise, the structural nested model must only has less number of parameters.

U_i: cluster-level unmeasured confounders;

X_{ij}: individual-level confounders;

$Y_{ij}(a)$: potential outcome to adherence with $A_i = a$;

When the cluster level of adherence variable A_i is one dimensional (dichotomous), a structural nested model involves potential outcomes and the simple linear model is

$$E(Y_{ij}(a)|A_i = a, Z_i) = E(Y_{ij}(0)|A_i = a, Z_i) + a\xi, \tag{4–2}$$

where $E(Y_{ij}(a) - Y_{ij}(0)|A_i = a) = a\xi$, especially when $a = 1$, is the risk difference among individuals within observed adherence equal to a.

Here we assume that the cluster level randomization Z_i has no direct effect on individual binary outcome Y_{ij} except through cluster level adherence A_i, which is also the exclusion assumption considered by Angrist et al. [4] (refer Figure 4-1).

The linear form structural nested model works well for some cases of dichotomous A_i; we can use linear structural nested models even with binary outcomes, provided that the model is saturated as a function of adherence. But for some choices of A_i, model (4–2) might be poorly specified, and we will turn to use logistic or logarithmic structural nested models. In the SWASH study, A_i is ordinal, then the model might predict that $E(Y_{ij}(0)|A_i = a)$ is negative or greater than one, which is undesirable with binary outcome Y_{ij}. In this case, we appeal to a link function as a solution to this problem. Here in this case we considered the logit link instead of the linear link in model (4–2).

Before set up the logit structural nested model, we denote ξ as the effect of adherences, and let $\xi_0 = E(Y_{ij}(0))$. Also let $\mu(A_i; Z_i; \eta)$ be a parametric model for $E(Y_{ij}|A_i; Z_i)$ with parameter η; here we let $\mu(A_i; Z_i; \eta) = h^{-1}(A_i\eta_1 + Z_i\eta_2 + A_i * Z_i\eta_3)$, where the link is $h(p) = log(\frac{p}{1-p})$ the logit transform and $A_i * Z_i$ represents a multidimensional interaction.

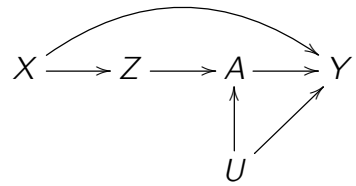

Figure 4-2. DAG represents the causal relationship between variables (with individual level confounders).

Then the structural nested model could be set up as

$$h(E(Y_{ij}(a)|A_i = a; Z_i)) = h(E(Y_{ij}(0)|A_i = a; Z_i)) + a\xi; \tag{4–3}$$

where ξ represent the effects of adherences expressed as the log odds ratios.

4.3 Methods

Here in this chapter, we would like to show how to adapt the structural nested model to instrumental-variable setting with complex survey data.

4.3.1 Estimation with complex sampling designs

In the SWASH study, we consider the individual level covariates X_{ij}, which may have confounding effect on estimating the effect of intervention arm reception on the binary outcome. Therefore we construct confounding weights w_{ij}^c to adjust the effect of individual level confounders. Assume that the treatment arm is independent to the potential baseline outcome given individual level confounders X_{ij}, i.e. $Z_i \perp\!\!\!\perp Y_{ij}(0)|X_{ij}$. So X_{ij} form a sufficient set of confounders. Then the confounding weights w_{ij}^c is estimated as the inverse of the probability that individual j receives treatment Z_i given the individual level confounders X_{ij}. Therefore we can use a new DAG (see Figure 4-2) include individual level confounders X_{ij} to represent the causal relationship.

Then the three IV assumptions regarding the new DAG can be expressed as

1. Z is independent of U;
2. Z is associated with A;
3. Z is independent of Y given U and X and A.

For model (4–3), we can then estimate ξ by solving the following estimating equations

$$\sum_i \sum_j w_{ij} Z_i^T (h^{-1}(h(\mu(A_i, Z_i; \eta)) - A_i\xi) - \xi_0) = 0, \tag{4–4}$$

$$\sum_i \sum_j w_{ij} D_i^T (Y_{ij} - \mu(A_i; Z_i; \eta)) = 0, \tag{4–5}$$

where D_i^T is a function of A_i and Z_i, e.g. $(A_i; Z_i; A_i * Z_i)^T$, and w_{ij} is the double weight for individual j within cluster i, which is product of the complex sampling weight and the confounding weight w_{ij}^c.

Assume that the w_{ij}, the structural nested model and $\mu(\cdot)$ is correctly specified, then these estimating equations generate a consistent estimator of ξ, because $Y_{ij}(0)$ is independent of Z_i assuming ideal (no confounding and equal probability) sampling, and

$E_{A_i|Z_i}\left(h^{-1}(h(E(Y_{ij}(a)|A_i = a, Z_i)) - A_i\xi)\right) = E(Y_{ij}(0)|Z_i) = \xi_0$.

Using Taylor expansion, the expit link h^{-1} can be linearized as

$$\begin{aligned} h^{-1}(l_i - A_i\xi^{(n+1)}) &\approx h^{-1}(l_i - A_i\xi^{(n)}) + \frac{\partial h^{-1}(l_i - A_i\xi)}{\partial \xi}|_{\xi=\xi^{(n)}}(\xi^{(n+1)} - \xi^{(n)}) \\ &= Y_i^* - A_i^*\xi^{(n+1)} \end{aligned}$$

where $Y_i^* = expit(l_i - A_i\xi^{(n)}) + \frac{exp(l_i - A_i\xi^{(n)})}{(1+exp(l_i - A_i\xi^{(n)}))^2} A_i\xi^{(n)}$ and $A_i^* = \frac{exp(l_i - A_i\xi^{(n)})}{(1+exp(l_i - A_i\xi^{(n)}))^2} A_i$. After linearization, equation (4–4) can be formed as

$$\sum_i \sum_j w_{ij} Z_i^T (Y_i^* - A_i^*\xi - \xi_0) = 0.$$

Then with updated ξ and ξ_0, the equation (4–4) can be solved using iterative standard weighted instrumental variables(IV) estimation. Equation (4–5) can be solved using weighted generalized linear model regression.

Next, we estimate $E(Y_{ij}(a)|A_i = a, Z_i)$ by the mean of individual outcome Y_{ij} conditionally on Z_i and A_i, which is

$$\hat{Y}_{\{A_i=a,Z_i=z\}}(a) = \frac{1}{\sum_i \sum_j w_{ij} I\{A_i = a, Z_i = z\}} \sum_i \sum_j w_{ij} Y_{ij} I\{A_i = a, Z_i = z\}.$$

Then we derive the estimator of $E(Y_{ij}(0)|A_i = a, Z_i)$ using the estimated $\hat{\xi}$, the structural nested model and the estimator of $E(Y_{ij}(a)|A_i = a, Z_i)$, i.e.

$$\hat{Y}_{ij}(0) = h^{-1}(h(Y_{ij}) - a\xi)$$

and

$$\hat{Y}_{\{A_i=a,Z_i=z\}}(0) = \frac{1}{\sum_i \sum_j w_{ij} I\{A_i = a, Z_i = z\}} \sum_i \sum_j w_{ij} \hat{Y}_{ij}(0) I\{A_i = a, Z_i = z\}.$$

We use the relationship $E(Y_{ij}(a)|A_i = a) = E_{Z_i|A_i} E(Y_{ij}(a)|A_i = a, Z_i)$ and $E(Y_{ij}(0)|A_i = a) = E_{Z_i|A_i} E(Y_{ij}(0)|A_i = a, Z_i)$ to estimate the relative risk (RR) and relative difference (RD) for each level of A_i i.e.

$$RR(A_i = a) = \frac{E(Y_{ij}(a)|A_i = a)}{E(Y_{ij}(0)|A_i = a)}, \tag{4–6}$$

$$RD(A_i = a) = E(Y_{ij}(a)|A_i = a) - E(Y_{ij}(0)|A_i = a). \tag{4–7}$$

Please refer the detailed calculation in Appendix D.

4.3.2 Variance Estimation

The method above mainly focus on solving an unbiased estimating equation of form $U(\theta) = \sum_{p=1}^{P} \sum_{c=1}^{C_p} U_{pc}(\theta) = 0$, where θ is a vector of parameter of interest, c indexes groups that play the role of primary sampling units (in SWASH study, we refer c as schools), p indexes the primary strata (in the SWASH study, we refer p as the two district areas), and $U_{pc}(\theta)$ is a sum of weighted estimating equations with the weighted components

each having a zero expectation. Let $\nabla U(\hat{\theta})$ be the gradient of $U(\hat{\theta})$, and let

$$\Omega = \sum_{c=1}^{C_p} (U_{pc}(\hat{\theta}) - U_{p\cdot}(\hat{\theta}))(U_{pc}(\hat{\theta}) - U_{p\cdot}(\hat{\theta}))^T,$$

$$V(\hat{\theta}) = \sum_{p=1}^{P} (C_p/(C_p - 1))\Omega,$$

where $U_{p\cdot}(\hat{\theta})$ is the mean of $\{U_{p1}(\hat{\theta}), \cdots, U_{pC_p}(\hat{\theta})\}$. Then we use sandwich estimator based on Taylor series (Binder [9]) as estimation of the variance $Var(\theta)$

$$\hat{Var}(\hat{\theta}) = [\nabla U(\hat{\theta})]^{-1} V(\hat{\theta}) [\nabla U(\hat{\theta})^T]^{-1}. \tag{4–8}$$

With the estimated variance of the parameters $\hat{Var}(\hat{\theta})$, we can estimate the variance of the relative risk $\hat{Var}(RR(A_i = 1))$ and $\hat{Var}(RR(A_i = 1))$ through Delta method [17]. Then the confidence intervals for the relative risks would be easy to estimate.

Please refer the detailed variance estimation calculation in Appendix E.

4.4 Simulation Study

In this section, we will discuss the validity of IV on structural nested model.

4.4.1 Simulation with strong instrument

Suppose the size of the dataset is 10,000. Let Z be a 3-category variable with equal probabilities of value 0, 1 and 2. Then we generate another 3-category variable A depend on Z with unequal probabilities. That is

For $Z = 0$,

Value of A	Probability
0	3/4
1	1/8
2	1/8

For $Z = 1$,

Value of A	Probability
0	1/8
1	3/4
2	1/8

For $Z = 2$,

Value of A	Probability
0	1/8
1	1/8
2	3/4

Then the structural nested model can be set up as the model below

$$logit(E(Y(a)|A = a, Z)) = logit(E(Y(0)|A = a, Z)) + a\xi,$$

where we set $\xi = \log 2$. The probabilities of $Y(0)$ can be set up as the 3×3 table,

Table 4-1. Distribution of $Y(0)|A = a, Z$.

P(Y(0)|A=a,Z)	$Z = 0$	$Z = 1$	$Z = 2$
$A = 0$	1/5	1/4	1/3
$A = 1$	1/4	1/5	1/4
$A = 0$	1/3	1/3	1/5

According the distribution in Table 4-1, we can prove that $Y(0) \amalg Z$, i.e.

$$\begin{aligned}
E(Y(0)|Z = 0) &= \frac{1}{5} \times \frac{3}{4} + \frac{1}{4} \times \frac{1}{8} + \frac{1}{3} + \frac{1}{8} = 0.6396, \\
E(Y(0)|Z = 1) &= \frac{1}{4} \times \frac{1}{8} + \frac{1}{5} \times \frac{3}{4} + \frac{1}{3} + \frac{1}{8} = 0.6396, \\
E(Y(0)|Z = 2) &= \frac{1}{3} \times \frac{1}{8} + \frac{1}{4} \times \frac{1}{8} + \frac{1}{5} + \frac{3}{4} = 0.6396.
\end{aligned}$$

Therefore $E(Y(0)|Z = 0) = E(Y(0)|Z = 1) = E(Y(0)|Z = 2)$. Then, using the structural nested model, the 3×3 table for the probability of $Y(a)$ is

Table 4-2. Distribution of $Y(a)|A=a, Z$.

P(Y(a)|A=a,Z)	$Z=0$	$Z=1$	$Z=2$
$A=0$	1/5	1/4	1/3
$A=1$	2/5	1/3	2/5
$A=2$	2/3	2/3	1/2

Therefore we can generate the binary outcome Y according to Table 4-2 and then calculate the conditional expectations by Table 4-1 and Table 4-2.

$$\begin{aligned}
E(Y(0)|A=1) &= \frac{1}{4}\times\frac{1}{8}+\frac{1}{5}\times\frac{3}{4}+\frac{1}{4}\times\frac{1}{8}=0.2125\\
E(Y(0)|A=2) &= \frac{1}{3}\times\frac{1}{8}+\frac{1}{3}\times\frac{1}{8}+\frac{1}{5}\times\frac{3}{4}=0.2333\\
E(Y(1)|A=1) &= \frac{2}{5}\times\frac{1}{8}+\frac{1}{3}\times\frac{3}{4}+\frac{2}{5}\times\frac{1}{8}=0.35\\
E(Y(2)|A=2) &= \frac{2}{3}\times\frac{1}{8}+\frac{2}{3}\times\frac{1}{8}+\frac{1}{2}\times\frac{3}{4}=0.5417
\end{aligned}$$

Therefore the RRs and RDs for each level of A are

$$\begin{aligned}
RR(A=1) &= \frac{E(Y(1)|A=1)}{E(Y(0)|A=1)}\approx 1.647\\
RR(A=2) &= \frac{E(Y(2)|A=2)}{E(Y(0)|A=2)}\approx 2.321\\
RD(A=1) &= E(Y(1)|A=1)-E(Y(0)|A=1)=0.1375\\
RD(A=2) &= E(Y(2)|A=2)-E(Y(0)|A=2)=0.3083
\end{aligned}$$

We repeat the simulation 100 times. The mean of the estimated RRs and RDs are shown in Table 4-3.

Table 4-3. Simulation results for IV method on a structural nested model with strong instrument.

	RR(A=1)	RR(A=2)	RD(A=1)	RD(A=2)
True value	1.647	2.321	0.1375	0.3083
IV estimator	1.657	2.331	0.1379	0.3079

Comparing the simulation results (see Table 4-3), the estimators are quite close to the true values. We can see that IV on the structural nested model with complex survey data performs very well.

4.4.2 Simulation with weak instrument

Similar as Section 4.4.1, suppose the size of the dataset is 10,000. Let Z be a 3-category variable with equal probabilities of value 0, 1 and 2. Then we generate another 3-category variable A depend on Z with unequal probabilities. That is

For $Z = 0$,

Value of A	Probability
0	1/6
1	1/3
2	1/2

For $Z = 1$,

Value of A	Probability
0	1/2
1	1/6
2	1/3

For $Z = 2$,

Value of A	Probability
0	1/3
1	1/2
2	1/6

The probabilities of $Y(0)$ and $Y(a)$ can be set up as the same ones in Section 4.4.1(See Table 4-1 and Table 4-2). Similarly we can prove $Y(0) \amalg Z$, i.e.

$$\begin{aligned}
E(Y(0)|Z=0) &= \frac{1}{5}\times\frac{1}{6}+\frac{1}{4}\times\frac{1}{3}+\frac{1}{3}+\frac{1}{2}=0.2833,\\
E(Y(0)|Z=1) &= \frac{1}{4}\times\frac{1}{3}+\frac{1}{5}\times\frac{1}{6}+\frac{1}{3}+\frac{1}{2}=0.2833,\\
E(Y(0)|Z=2) &= \frac{1}{3}\times\frac{1}{2}+\frac{1}{4}\times\frac{1}{3}+\frac{1}{5}+\frac{1}{6}=0.2833.
\end{aligned}$$

Therefore $E(Y(0)|Z=0) = E(Y(0)|Z=1) = E(Y(0)|Z=2)$.

Therefore we can generate the binary outcome Y according to Table 4-2 and then calculate the conditional expectations by Table 4-1 and Table 4-2.

$$
\begin{aligned}
E(Y(0)|A=1) &= \frac{1}{4}\times\frac{1}{3}+\frac{1}{5}\times\frac{1}{6}+\frac{1}{4}\times\frac{1}{2}=0.2417\\
E(Y(0)|A=2) &= \frac{1}{3}\times\frac{1}{2}+\frac{1}{3}\times\frac{1}{3}+\frac{1}{5}\times\frac{1}{6}=0.3111\\
E(Y(1)|A=1) &= \frac{2}{5}\times\frac{1}{3}+\frac{1}{3}\times\frac{1}{6}+\frac{2}{5}\times\frac{1}{2}=0.3889\\
E(Y(2)|A=2) &= \frac{2}{3}\times\frac{1}{2}+\frac{2}{3}\times\frac{1}{3}+\frac{1}{2}\times\frac{1}{6}=0.6390
\end{aligned}
$$

Therefore the RRs and RDs for each level of A are

$$
\begin{aligned}
RR(A=1) &= \frac{E(Y(1)|A=1)}{E(Y(0)|A=1)}\approx 1.609\\
RR(A=2) &= \frac{E(Y(2)|A=2)}{E(Y(0)|A=2)}\approx 2.054\\
RD(A=1) &= E(Y(1)|A=1)-E(Y(0)|A=1)\approx 0.147\\
RD(A=2) &= E(Y(2)|A=2)-E(Y(0)|A=2)\approx 0.328
\end{aligned}
$$

We repeat the simulation 100 times. The mean of the estimated RRs and RDs are shown in Table 4-4.

Table 4-4. Simulation results for IV method on a structural nested model with weak instrument.

	RR(A=1)	RR(A=2)	RD(A=1)	RD(A=2)
True value	1.609	2.054	0.147	0.328
IV estimator	1.831	2.454	0.172	0.376

Comparing the simulation results (see Table 4-4), we can see that the biases are relatively small. We conclude that IV on the structural nested model with weak instrument also performs well.

4.5 Results on School-based Water, Sanitation, and Hygiene (SWASH) Project

First we use the notations below to index the variables used for analysis in the study.

i: schools in the SWASH study, $i = 1, \cdots, 86$;

j: individual pupil within the school i;

Y_{ij}: pupil absenteeism, which is an binary individual-level outcome;

Z_i: intervention arms, which is cluster level randomization. We use 0 for the control group , 1 for HP & WT group, 2 for HP & WT + sanitation group;

A_i: adherence to the intervention, which considered as cluster level adherence. Here the adherence is not assigned according to the Z_i. Since some schools were observed with safe water but no soap, we simply use 0 refers the schools had lowest levels of safe water, soap and sanitation (better latrine construction), 1 refers the group that have only one treatment, safe water or soap or latrine construction, while 2 refers the group that have at least two treatments; how to construct A_i will be represented in the next paragraph;

U_i: cluster-level unmeasured confounders;

X_{ij}: individual-level confounders, including educational level, age, gender;

w_{ij}: the double weight for individual j within cluster i, which is product of the complex sampling weight and the confounding weight. The confounding weight is a weight representing the estimated inverse probability of treatment as a function of individual-level covariates.

Note that Z_i and A_i are three level category variables. We generate three new binary variables C_1, C_2 and C_3, with C_1 representing whether soap was supplied on the day of measurement, C_2 representing whether the school had a satisfactory ratio of latrines to students on the day of measurement and C_3 indicating if there was safe water supplied on the day of measurement. Next we categorize the cluster-level adherence into three levels represented by A_i. Set $A_i = 0$ if $C_1 + C_2 + C_3 = 0$, $A_i = 1$ if $C_1 + C_2 + C_3 = 1$, and $A_i = 2$ if $C_1 + C_2 + C_3 = 2$ or 3. We then generate two binary variables corresponding to A_i, i.e. $A_{i1} = 1$ if $A_i = 1$ and $A_{i2} = 1$ if $A_i = 2$. Similarly, set $Z_{i1} = 1$ if $Z_i = 1$ and $Z_{i2} = 1$ if $Z_i = 2$. We model the adherence effects ξ using two dimensions, i.e. $\xi = (\xi_1, \xi_2)^T$. Therefore we change the multi-dimension variables Z_i and A_i to a group of multi variables (Z_{i1}, Z_{i2}, A_{i1} and A_{i1}) with one dimension.

Now we apply the IV method with the logit form structural nested model to the SWASH study. We estimate ξ in the structural nested model by solving the estimating

equations using SAS software (Version 9.2). The estimates are $\hat{\theta}_1 = 1.3474$ and $\hat{\theta}_2 = 0.4749$. The estimated relative risks are $\hat{RR}(A_i = 1) = 0.4086$ and $\hat{RR}(A_i = 2) = 0.6938$. The corresponding 95% confidence intervals are (0.19, 0.86) and (0.41, 1.17).

4.6 Conclusions

We developed and implemented a methodology to estimate the effect of school-level adherence on individual absenteeism in the SWASH trial in Western Kenya. Specifically, we extended instrumental variable (IV) methods based on structural nested models to accommodate data from complex sampling designs. We also considered ordinal compliance in a three-armed trial; most previous research considers binary compliance and two-armed trials. Our results indicate that the effect of adopting exactly one of the three provisions (soap, sanitation, or safe water) is to decrease the risk of absenteeism by 41% versus adopting no provision, within the group of schools that adopted exacting one provision; with 95% confidence level, the risk lies in the confidence interval $(19\%, 86\%)$. Further, the effect of adopting two or more of the three provisions is to decrease the risk of absenteeism by 70% versus adopting no provision, within the group of schools that adopted two or three provisions; with 95% confidence level, the risk lies in the confidence interval $(41\%, 117\%)$. Thus, our results suggest that the provisions are effective in reducing absenteeism.

APPENDIX A
PROOF OF LEMMA 1 (CHAPTER 2)

Lemma 1. For $C_i \neq 0_{p\times 1}$,

$$\lim_{k\to\infty} \frac{\sum_{r=0}^{k} \binom{k}{r}^2 (k-r) e^{(k-r)C_i^T\beta}}{\sum_{r=0}^{k} \binom{k}{r}^2 k e^{(k-r)C_i^T\beta}} = \frac{e^{C_i^T\frac{\beta}{2}}}{1+e^{C_i^T\frac{\beta}{2}}}. \tag{A–1}$$

Proof. Let $y = C_i^T\beta$ and $g_k(y) = \sum_{r=0}^{k} \binom{k}{r}^2 e^{(k-r)y}$. The left hand side of (A–1) may be expressed as

$$\lim_{k\to\infty} \frac{\sum_{r=0}^{k} \binom{k}{r}^2 (k-r) e^{(k-r)y}}{\sum_{r=0}^{k} \binom{k}{r}^2 k e^{(k-r)y}} = \lim_{k\to\infty} \frac{g_k'(y)}{k g_k(y)}. \tag{A–2}$$

Now let $x = (1+e^y)/(1-e^y)$. Clearly, $|x| > 1$ for any y and

$$g_k(y) = (1+x)^{-k} \sum_{r=0}^{k} \binom{k}{r}^2 k(x-1)^{(k-r)}(x+1)^r = 2^k(1+x)^{-k}P_k(x), \tag{A–3}$$

and

$$g_k'(y) = 2^k(1+x)^{-k}\Big\{\frac{-k}{1+x}P_k(x) + P_k'(x)\Big\}\frac{dx}{dy}, \tag{A–4}$$

where $P_k(x) = 2^{-k}\sum_{r=0}^{k} \binom{k}{r}^2 k(x-1)^{(k-r)}(x+1)^r$ is the Legendre polynomial of degree k. Hence we have

The Legendre polynomial $P_k(x)$ has the following well known properties for all $|x| > 1$:

(a) $P_k^2(x) - P_{k+1}(x)P_{k-1}(x) \leq 0$,

(b) $(k+1)P_{k+1}(x) - (2k+1)xP_k(x) + kP_{k-1}(x) = 0$ (Bonnet's recursion formula),

(c) $(1-x^2)P_k'(x) = (k+1)\{xP_k(x) - P_{k+1}(x)\}$.

Since $P_k(x) > 0$ for $x > 1$ and $P_k(x) = (-1)^k P_k(-x)$ for $x < -1$, we may easily see from (a) that $P_{k+1}(x)/P_k(x) \geq P_k(x)/P_{k-1}(x) > 0$, for $x > 1$ and $P_{k+1}(x)/P_k(x) \leq P_k(x)/P_{k-1}(x) < 0$, for $x > -1$. From (b), we see that

$$\frac{P_{k+1}(x)}{P_k(x)} = \frac{2k+1}{k+1}x - \frac{k}{k+1}\frac{P_{k-1}(x)}{P_k(x)} \begin{cases} < 2x & \text{if } x > 1 \\ > 2x & \text{if } x < -1 \end{cases},$$

which means the sequence $P_{k+1}(x)/P_k(x)$ is a non-decreasing positive sequence with upper bound $2x$ if $x > 1$ and a non-increasing negative sequence with lower bound $2x$ if $x < -1$. Let $a_P = \lim_{k\to\infty} P_{k+1}(x)/P_k(x)$. Divide the equality (b) by $kP_k(x)$ and let $k \to \infty$, we have that

$$a_P - 2x + 1/a_P = 0 \quad \text{gives} \quad a_P = x \pm \sqrt{x^2 - 1}.$$

That is, the sequence $P_{k+1}(x)/P_k(x)$ converges to

$$\lim_{k\to\infty} \frac{P_{k+1}(x)}{P_k(x)} = \begin{cases} x + \sqrt{x^2-1} & \text{if } x > 1 \\ x - \sqrt{x^2-1} & \text{if } x < -1 \end{cases}.$$

As a result, the limitation in (A–2) may be written as

$$\begin{aligned}
\lim_{k\to\infty} \frac{g_k'(y)}{kg_k(y)} &= \lim_{k\to\infty} \frac{\left\{\frac{-k}{1+x} P_k(x) + P_k'(x)\right\}\frac{dx}{dy}}{kP_k(x)} \\
&= \left[\frac{-1}{1+x} + \frac{1}{x^2-1}\left\{\lim_{k\to\infty} \frac{P_{k+1}(x)}{P_k(x)} - x\right\}\right]\frac{x^2-1}{2} \\
&= \begin{cases} \frac{1}{2}(1 - x + \sqrt{x^2-1}) & \text{if } x > 1 \\ \frac{1}{2}(1 - x - \sqrt{x^2-1}) & \text{if } x < -1 \end{cases} \\
&= \frac{e^{y/2}}{1 + e^{y/2}} \\
&= \frac{e^{C_i^T\beta/2}}{1 + e^{C_i^T\beta/2}}.
\end{aligned}$$

APPENDIX B
PROOF OF LEMMA 2 (CHAPTER 2)

Lemma 2. *Without loss of generality, suppose $C_{is} > 0$, $s = 1, \cdots, p$. When $\beta_s \geq 0$, $\frac{\sum_{r=0}^{k}\binom{k}{r}^2(k-r)e^{(k-r)C_{is}\beta_s}}{\sum_{r=0}^{k}\binom{k}{r}^2 ke^{(k-r)C_{is}\beta_s}}$ is non-increasing with $k \to \infty$; when $\beta_s < 0$, $\frac{\sum_{r=0}^{k}\binom{k}{r}^2(k-r)e^{(k-r)C_{is}\beta_s}}{\sum_{r=0}^{k}\binom{k}{r}^2 ke^{(k-r)C_{is}\beta_s}}$ is non-decreasing with $k \to \infty$.*

Proof. From the argument in Appendix A, since

$$\frac{\sum_{r=0}^{k}\binom{k}{r}^2(k-r)e^{(k-r)C_i^T\beta}}{\sum_{r=0}^{k}\binom{k}{r}^2 ke^{(k-r)C_i^T\beta}} = \left[\frac{-1}{1+x} + \frac{k+1}{k(x^2-1)}\left\{\frac{P_{k+1}(x)}{P_k(x)} - x\right\}\right]\frac{x^2-1}{2}, \tag{B–1}$$

where $x = (1+e^{C_i^T\beta})/(1-e^{C_i^T\beta})$, the monotonicity of (B–1) is the same as the sequence

$$h_k(x) = \frac{k+1}{k}\left\{\frac{P_{k+1}(x)}{P_k(x)} - x\right\}.$$

After a few algebras, we have

$$\begin{aligned}
h_k(x) - h_{k+1}(x) &= \frac{k+1}{k}\left\{\frac{P_{k+1}(x)}{P_k(x)} - x\right\} - \frac{k+2}{(k+1)}\left\{\frac{P_{k+2}(x)}{P_{k+1}(x)} - x\right\} \\
&= \frac{k+1}{k}\left\{\frac{P_{k+1}(x)}{P_k(x)} - x\right\} \\
&\quad - \frac{k+2}{(k+1)}\left[\frac{\frac{1}{k+2}\{(2k+3)xP_{k+1}(x) - (k+1)P_k(x)\}}{P_{k+1}(x)} - x\right] \\
&= \frac{(k+1)P_{k+1}^2(x) - (2k+1)xP_k(x)P_{k+1}(x) + kP_k^2(x)}{kP_k(x)P_{k+1}(x)} \\
&= \frac{P_k^2(x) - P_{k+1}(x)P_{k-1}(x)}{P_k(x)P_{k+1}(x)},
\end{aligned}$$

where the second and third equalities are due to the property (b) in Appendix A. Now because of the inequality (c) in Appendix A and the fact that $P_k(x)P_{k+1}(x) > 0$ for $x > 1$ and $P_k(x)P_{k+1}(x) = (-1)^{2k+1}P_k(-x)P_{k+1}(-x) < 0$ for $x < -1$, we see that $h_k(x) - h_{k+1}(x) \geq 0$ for $x > 1$ and ≤ 0 for $x < -1$. That is, the sequence in (B–1) is non-decreasing if $x > 1$ or $C_i^T\beta < 0$ and non-increasing if $x < -1$ or $C_i^T\beta > 0$.

APPENDIX C
SAS CODE (CHAPTER 3)

All programming was done using SAS software (version 9.2). We used SAS PROC SQL to create the paired data set. Specifically,

```
proc sql;
    create table match as
    select
    one.hhx,
    one.wtia as wtiapair
    one.strat,
    one.psu,
    (one.educ - two.educ) as educd,
    (one.cover - two.cover) as coverd
    from nhis2009.personclean one, nhis2009.personclean two
    where (one.hhx=two.hhx and one.cover>two.cover);
```

In the above code, nhis2009.personclean is the NHIS data set that has been cleaned in the sense of deleting observations with missing data for education or coverage. The variable `hhx` identifies an individuals household. The variable `wtia` is the interim annual weight in the NHIS data, and it is identical for individuals within the same household. The variables `strat` and `psu` are the NHIS strata and PSU variables, and `educ` and `cover` are the variables we created for dichotomous education and health insurance coverage. The new data set with paired observations is called match, and it has variables `hhx`, `wtiapair`, `strat`, `psu`, `educd`, and `coverd`. The variable `educd` is the differenced education variable, whereas `coverd` is equal to one for everyone. To create some observations with outcomes equal to zero, we used the SAS code

```
data nhis2009.match;
    set match;
```

```
    if hhx < 1000 then do;
    coverd=0;
    educd=-educd;
    end;
run;
```

Finally, to implement the logistic regression, we used the SAS code

```
proc surveylogistic data=nhis2009.match;
    strata strat;
    cluster psu;
    model coverd (event=1)= educd / noint;
    weight wtiapair;
run;
```

APPENDIX D
ESTIMATION OF INSTRUMENTAL VARIABLE ON STRUCTURE NESTED MODELS

Suppose Z_i and A_i are 3-category variables in model (4–3). First we generate two new variables for each A_i and Z_i, i.e. A_{i1}, A_{i2}, Z_{i1} and Z_{i2}. They we generated as below

$Z_{i1} = 0$ and $Z_{i2} = 0$ for $Z_i = 0$;

$Z_{i1} = 1$ and $Z_{i2} = 0$ for $Z_i = 1$;

$Z_{i1} = 0$ and $Z_{i2} = 1$ for $Z_i = 2$;

$A_{i1} = 0$ and $A_{i2} = 0$ for $A_i = 0$;

$A_{i1} = 1$ and $A_{i2} = 0$ for $A_i = 1$;

$A_{i1} = 0$ and $A_{i2} = 1$ for $A_i = 2$.

$\mu(A_i, Z_i; \eta)$ is defined as a parametric model for $E(Y_{ij}|A_i, Z_i)$ with parameter η, where in Chapter 4 it is defined as

$$\begin{aligned} & \mu(A_i, Z_i; \eta) \qquad \text{(D–1)} \\ = \; & h^{-1}(A_i\eta_1 + Z_i\eta_2 + A_i * Z_i\eta_3) \\ = \; & expit(\eta_{10} + A_{i1}\eta_{11} + A_{i2}\eta_{12} + Z_{i1}\eta_{21} + Z_{i2}\eta_{22} + A_{i1}Z_{i1}\eta_{31} + A_{i1}Z_{i2}\eta_{32} + A_{i2}Z_{i1}\eta_{33} + A_{i2}Z_{i2}\eta_{34}), \end{aligned}$$

where *expit* means the inverse mapping of *logit*, i.e. $expit(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$. Here we define the linear sum in equation (D–1) as l_i, therefore $\mu(A_i, Z_i; \eta) = expit(l_i)$.

Assume the treatment arm is independent of the potential baseline outcome given individual confounder X_{ij}, i.e. $Z_i \amalg Y_{ij}(0)|X_{ij}$. With double weights ($w_{ij}$), the product of the sampling design weights and a weight representing the estimated inverse probability of treatment as a function of individual-level confounders, the potential baseline outcome is independent of treatment arm, $Z_i \amalg Y_{ij}(0)$.

We are interested in estimating relative risk for each level of A_i by estimating ξ, ξ_0, and η. We can estimate the parameters by solving the estimating equations

$$\sum_i \sum_j w_{ij} Z_i^T (h^{-1}(h(\mu(A_i, Z_i; \eta)) - A_i\xi) - \xi_0) = 0, \tag{D–2}$$

$$\sum_i \sum_j w_{ij} D_i^T (Y_{ij} - \mu(A_i, Z_i; \eta)) = 0, \tag{D–3}$$

where D_i^T is a function of A_i and Z_i, e.g. $(A_i, Z_i, A_i * Z_i)$. We also know that $\xi_0 = E(Y_{ij}(0)) = E(Y_{ij}(0)|Z_i) = E_{A_i|Z_i}[h^{-1}(h(E(Y_{ij}(a)|A_i = a, Z_i)) - A_i\xi)]$.

Write equation (D–3) in details,

$$\begin{aligned}
\sum_i \sum_j w_{ij}(Y_{ij} - expit(l_i)) &= 0, \\
\sum_i \sum_j w_{ij} A_{i1}(Y_{ij} - expit(l_i)) &= 0, \\
\sum_i \sum_j w_{ij} A_{i2}(Y_{ij} - expit(l_i)) &= 0, \\
\sum_i \sum_j w_{ij} Z_{i1}(Y_{ij} - expit(l_i)) &= 0, \\
\sum_i \sum_j w_{ij} Z_{i2}(Y_{ij} - expit(l_i)) &= 0, \\
\sum_i \sum_j w_{ij} A_{i1} Z_{i1}(Y_{ij} - expit(l_i)) &= 0, \\
\sum_i \sum_j w_{ij} A_{i1} Z_{i2}(Y_{ij} - expit(l_i)) &= 0, \\
\sum_i \sum_j w_{ij} A_{i2} Z_{i1}(Y_{ij} - expit(l_i)) &= 0, \\
\sum_i \sum_j w_{ij} A_{i2} Z_{i2}(Y_{ij} - expit(l_i)) &= 0.
\end{aligned} \tag{D–4}$$

Then we can estimate η using weighted generalized linear model (wGLM) regression method. We obtain the estimators using “surveylogistic” procedure in SAS software (Version 9.2).

Next we substitute the logit transform for h in equation (D–3), then we get

$$\sum_i \sum_j w_{ij} Z_i^T (expit(l_i - A_i\xi) - \xi_0) = 0. \tag{D–5}$$

Write the equation in detail, it turns out

$$\begin{aligned}
&\sum_i \sum_j w_{ij}(expit(l_i - A_{i1}\xi_1 - A_{i2}\xi_2) - \xi_0) = 0, \\
&\sum_i \sum_j w_{ij} Z_{i1}(expit(l_i - A_{i1}\xi_1 - A_{i2}\xi_2) - \xi_0) = 0, \\
&\sum_i \sum_j w_{ij} Z_{i2}(expit(l_i - A_{i1}\xi_1 - A_{i2}\xi_2) - \xi_0) = 0.
\end{aligned}$$

To linearise the *expit* part in the equations above, we define $f(\xi) = expit(l_i - A_i\xi)$. The gradient of $f(\xi)$ is $\frac{\partial f(\xi)}{\partial \xi} = \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} A_i$. Using Taylor expansion with the $(n-1)$th known $\xi^{(n-1)}$, we approximately have

$$\begin{aligned}
f(\xi_{(n)}) &\approx f(\xi_{(n-1)}) + \frac{\partial f(\xi)}{\partial \xi}|_{\xi=\xi^{(n-1)}}(\xi^{(n)} - \xi^{(n-1)}) \\
&= expit(l_i - A_i\xi^{(n-1)}) - \frac{\exp(l_i - A_i\xi^{(n-1)})}{[1+\exp(l_i - A_i\xi^{(n-1)})]^2} A_i(\xi^{(n)} - \xi^{(n-1)}) \\
&= y_i^* - A_i^*\xi^{(n)},
\end{aligned}$$

where

$$\begin{aligned}
y_i^* &= expit(l_i - A_i\xi^{(n-1)}) + \frac{\exp(l_i - A_i\xi^{(n-1)})}{[1+\exp(l_i - A_i\xi^{(n-1)})]^2} A_i\xi^{(n-1)} \\
&= \frac{m_i^*}{1+m_i^*} + \frac{m_i^*}{(1+m_i^*)^2} A_i\xi^{(n-1)} \\
A_i^* &= \frac{\exp(l_i - A_i\xi^{(n-1)})}{[1+\exp(l_i - A_i\xi^{(n-1)})]^2} A_i \\
&= \frac{m_i^*}{(1+m_i^*)^2} A_i \\
m_i^* &= \exp(l_i - A_i\xi^{(n-1)}).
\end{aligned}$$

After linearization, equation (D–5) then becomes

$$\sum_i \sum_j w_{ij} Z_i^T (y_i^* - A_i^*\xi - \xi_0) = 0. \tag{D–6}$$

In summary, the algorithm for estimating ξ is descried as below

1. Set initial values of $\xi_0^{(0)}$ and $\xi^{(0)}$; set $j = 0$, let $\xi_0^{(j)} = \xi_0^{(0)}$, $\xi^{(j)} = \xi^{(0)}$.

2. Obtain $m_i^{*(j)}$, $y_i^{*(j)}$ and $A_i^{*(j)}$ as

$$\begin{aligned} m_i^{*(j)} &= expit(l_i - A_i\xi^{(j)}), \\ y_i^{*(j)} &= \frac{m_i^{*(j)}}{1+m_i^{*(j)}} + \frac{m_i^{*(j)}}{(1+m_i^{*(j)})^2}A_i\xi^{(j)}, \\ A_i^{*(j)} &= \frac{m_i^{*(j)}}{(1+m_i^{*(j)})^2}A_i. \end{aligned}$$

3. Solve Equation (D–6) with $y_i^{*(j)}$ and $A_i^{*(j)}$ to obtain $\xi_0^{(j+1)}$ and $\xi^{(j+1)}$.

4. For some criteria a, if $\|\xi_0^{(j+1)} - \xi_0^{(j)}\| + \|\xi^{(j+1)} - \xi^{(j)}\| < a$, then stop and return with $\hat{\xi} = \xi^{(j+1)}$ and $\hat{\xi}_0 = \xi_0^{(j+1)}$; otherwise set $j = j + 1$ and repeat step 2.

After estimating ξ, we can estimate relative risks (RR) for each level of adherence A_i, i.e.

$$RR(A_i = a) = \frac{E(Y_{ij}(a)|A_i = a)}{E(Y_{ij}(0)|A_i = a)},$$
$$RD(A_i = a) = E(Y_{ij}(a)|A_i = a) - E(Y_{ij}(0)|A_i = a).$$

First we estimate $E(Y_{ij}(a)|A_i = a, Z_i = z)$ by the weighted mean of individual outcome Y_{ij} conditionally on $A_i = a$ and $Z_i = z$, i.e. $\frac{\sum_i \sum_j w_{ij} Y_{ij} I\{A_i=a, Z_i=z\}}{\sum_i \sum_j w_{ij} I\{A_i=a, Z_i=z\}}$, and $E(Y_{ij}(a)|A_i = a)$ by weighted mean of individual outcome conditionally on $A_i = a$, i.e. $\frac{\sum_i \sum_j w_{ij} Y_{ij} I\{A_i=a\}}{\sum_i \sum_j w_{ij} I\{A_i=a\}}$.

Considering the structural nested model, we then have

$$E(Y_{ij}(0)|A_i = a, Z_i) = h^{-1}(h(E(Y_{ij}(a)|A_i = a, Z_i)) - a\xi), \tag{D–7}$$

we will obtain the estimate of $E(Y_{ij}(0)|A_i = a, Z_i)$ by substituting the estimate of $E(Y_{ij}(0)|A_i = a, Z_i)$ into equation (D–7). Define $q(A_i = a, Z_i)$ as the estimate of $E(Y_{ij}(0)|A_i = a, Z_i)$, then take expectation of Z_i given A_i on both side of equation (D–7), we have

$$E(Y_{ij}(0)|A_i = a) = E_{Z_i|A_i}(h^{-1}h(E(Y_{ij}(a)|A_i = a, Z_i)) - A_i\xi).$$

We then obtain the estimate of $E(Y_{ij}(0)|A_i)$, i.e.

$$\frac{\sum_{\textbf{all possible z}}(\sum_i \sum_j w_{ij} I\{A_i = a, Z_i = z\} q(A_i = a, Z_i = z))}{\sum_i \sum_j w_{ij} I\{A_i = a\}}.$$

Finally we can obtain $RR(A_i = a)$ by the derived estimators of $E(Y_{ij}(a)|A_i = a)$ and $E(Y_{ij}(0)|A_i = a)$.

APPENDIX E
VARIANCE ESTIMATION

Now we focus our interest in estimating variances of $RR(A_i = 1)$, $RR(A_i = 2)$, $RD(A_i = 1)$ and $RD(A_i = 2)$ by estimating the variances of $\hat{\xi}$, $\hat{\xi}_0$ and $\hat{\eta}$.

The method we developed here mainly focus on solving an unbiased estimating equation of form $U(\theta) = \sum_{p=1}^{P} \sum_{c=1}^{C_p} U_{pc}(\theta) = 0$, where θ is a vector of interest, c indexes groups that play the role of primary sampling units, p indexes the primary strata, $U_{pc}(\theta)$ is a sum of weighted estimating equations with the weighted components each having a zero expectation. Let $\nabla U(\hat{\theta})$ be the gradient of $U(\hat{\theta})$, and let

$$V(\hat{\theta}) = \sum_{p=1}^{P} \frac{C_p}{C_p - 1}(U_{pc}(\hat{\theta}) - U_{p\cdot}(\hat{\theta})(U_{pc}(\hat{\theta}) - U_{p\cdot}(\hat{\theta})^T,$$

where $U_{p\cdot}(\hat{\theta})$ is the mean of $\{U_{p1}(\hat{\theta}), U_{p2}(\hat{\theta}), \cdots, U_{pC_p}(\hat{\theta})\}$. then we use sandwich estimator based on Taylor series as estimation of the variance $Var(\theta)$,

$$\hat{Var}(\hat{\theta}) = [\nabla U(\hat{\theta})]^{-1} V(\hat{\theta})[\nabla U(\hat{\theta})^T]^{-1}. \tag{E–1}$$

Therefore, for equation (D–5), let

$$\begin{aligned}
U_{a1} &= \sum_i \sum_j w_{ij}(expit(l_i - A_{i1}\xi_1 - A_{i2}\xi_2) - \xi_0) = 0, \\
U_{a2} &= \sum_i \sum_j w_{ij} Z_{i1}(expit(l_i - A_{i1}\xi_1 - A_{i2}\xi_2) - \xi_0) = 0, \\
U_{a3} &= \sum_i \sum_j w_{ij} Z_{i2}(expit(l_i - A_{i1}\xi_1 - A_{i2}\xi_2) - \xi_0) = 0.
\end{aligned}$$

For Equation (D–4), let

$$
\begin{aligned}
U_{b1} &= \sum_i \sum_j w_{ij}(Y_{ij} - expit(l_i)) = 0, \\
U_{b2} &= \sum_i \sum_j w_{ij} A_{i1}(Y_{ij} - expit(l_i)) = 0, \\
U_{b3} &= \sum_i \sum_j w_{ij} A_{i2}(Y_{ij} - expit(l_i)) = 0, \\
U_{b4} &= \sum_i \sum_j w_{ij} Z_{i1}(Y_{ij} - expit(l_i)) = 0, \\
U_{b5} &= \sum_i \sum_j w_{ij} Z_{i2}(Y_{ij} - expit(l_i)) = 0, \\
U_{b6} &= \sum_i \sum_j w_{ij} A_{i1} Z_{i1}(Y_{ij} - expit(l_i)) = 0, \\
U_{b7} &= \sum_i \sum_j w_{ij} A_{i1} Z_{i2}(Y_{ij} - expit(l_i)) = 0, \\
U_{b8} &= \sum_i \sum_j w_{ij} A_{i2} Z_{i1}(Y_{ij} - expit(l_i)) = 0, \\
U_{b9} &= \sum_i \sum_j w_{ij} A_{i2} Z_{i2}(Y_{ij} - expit(l_i)) = 0.
\end{aligned}
$$

Let $v_i = \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2}$. Then we take partial derivatives of $U_{a1}, \cdots, U_{a3}, U_{b1}, \cdots, U_{b9}$ with respect to $\xi_0, \xi_1, \xi_2, \eta_{10}, \cdots, \eta_{34}$.

For ξ_0,

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \xi_0} &= -\sum_i \sum_j w_{ij}, \\
\frac{\partial U_{a2}}{\partial \xi_0} &= -\sum_i \sum_j w_{ij} Z_{i1}, \\
\frac{\partial U_{a3}}{\partial \xi_0} &= -\sum_i \sum_j w_{ij} Z_{i2}, \\
\frac{\partial U_{b1}}{\partial \xi_0} &= \cdots = \frac{\partial U_{b9}}{\partial \xi_0} = 0.
\end{aligned}
$$

For ξ_1,

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \xi_1} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i1} = -\sum_i \sum_j w_{ij} v_i A_{i1}, \\
\frac{\partial U_{a2}}{\partial \xi_1} &= -\sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i1} = -\sum_i \sum_j w_{ij} v_i Z_{i1} A_{i1}, \\
\frac{\partial U_{a3}}{\partial \xi_1} &= -\sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i1} = -\sum_i \sum_j w_{ij} v_i Z_{i2} A_{i1}, \\
\frac{\partial U_{b1}}{\partial \xi_1} &= \cdots = \frac{\partial U_{b9}}{\partial \xi_1} = 0.
\end{aligned}
$$

For ξ_2,

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \xi_2} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i2} = -\sum_i \sum_j w_{ij} v_i A_{i2}, \\
\frac{\partial U_{a2}}{\partial \xi_2} &= -\sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i2} = -\sum_i \sum_j w_{ij} v_i Z_{i1} A_{i2}, \\
\frac{\partial U_{a3}}{\partial \xi_2} &= -\sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i2} = -\sum_i \sum_j w_{ij} v_i Z_{i2} A_{i2}, \\
\frac{\partial U_{b1}}{\partial \xi_2} &= \cdots = \frac{\partial U_{b9}}{\partial \xi_2} = 0.
\end{aligned}
$$

For η_{10},

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \eta_{10}} &= \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} = \sum_i \sum_j w_{ij} v_i, \\
\frac{\partial U_{a2}}{\partial \eta_{10}} &= \sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} = \sum_i \sum_j w_{ij} v_i Z_{i1}, \\
\frac{\partial U_{a3}}{\partial \eta_{10}} &= \sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} = \sum_i \sum_j w_{ij} v_i Z_{i2}, \\
\frac{\partial U_{b1}}{\partial \eta_{10}} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} = -\sum_i \sum_j w_{ij} m_i, \\
\frac{\partial U_{b2}}{\partial \eta_{10}} &= -\sum_i \sum_j w_{ij} A_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} = -\sum_i \sum_j w_{ij} m_i A_{i1}, \\
\frac{\partial U_{b3}}{\partial \eta_{10}} &= -\sum_i \sum_j w_{ij} A_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} = -\sum_i \sum_j w_{ij} m_i A_{i2}, \\
\frac{\partial U_{b4}}{\partial \eta_{10}} &= -\sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} = -\sum_i \sum_j w_{ij} m_i Z_{i1}, \\
\frac{\partial U_{b5}}{\partial \eta_{10}} &= -\sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} = -\sum_i \sum_j w_{ij} m_i Z_{i2}, \\
\frac{\partial U_{b6}}{\partial \eta_{10}} &= -\sum_i \sum_j w_{ij} A_{i1} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i1}, \\
\frac{\partial U_{b7}}{\partial \eta_{10}} &= -\sum_i \sum_j w_{ij} A_{i1} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i2}, \\
\frac{\partial U_{b8}}{\partial \eta_{10}} &= -\sum_i \sum_j w_{ij} A_{i2} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i1}, \\
\frac{\partial U_{b9}}{\partial \eta_{10}} &= -\sum_i \sum_j w_{ij} A_{i2} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i2}.
\end{aligned}
$$

For η_{11},

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \eta_{11}} &= \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i1} = \sum_i \sum_j w_{ij} v_i A_{i1}, \\
\frac{\partial U_{a2}}{\partial \eta_{11}} &= \sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i1} = \sum_i \sum_j w_{ij} v_i Z_{i1} A_{i1}, \\
\frac{\partial U_{a3}}{\partial \eta_{11}} &= \sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i1} = \sum_i \sum_j w_{ij} v_i Z_{i2} A_{i1}, \\
\frac{\partial U_{b1}}{\partial \eta_{11}} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1}, \\
\frac{\partial U_{b2}}{\partial \eta_{11}} &= -\sum_i \sum_j w_{ij} A_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1}, \\
\frac{\partial U_{b3}}{\partial \eta_{11}} &= -\sum_i \sum_j w_{ij} A_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} = 0, (\because A_{i1} A_{i2} = 0), \\
\frac{\partial U_{b4}}{\partial \eta_{11}} &= -\sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i1}, \\
\frac{\partial U_{b5}}{\partial \eta_{11}} &= -\sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i2}, \\
\frac{\partial U_{b6}}{\partial \eta_{11}} &= -\sum_i \sum_j w_{ij} A_{i1} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i1}, \\
\frac{\partial U_{b7}}{\partial \eta_{11}} &= -\sum_i \sum_j w_{ij} A_{i1} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i1}, \\
\frac{\partial U_{b8}}{\partial \eta_{11}} &= 0, \\
\frac{\partial U_{b9}}{\partial \eta_{11}} &= 0.
\end{aligned}
$$

For η_{12},

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \eta_{12}} &= \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i2} = \sum_i \sum_j w_{ij} v_i A_{i2}, \\
\frac{\partial U_{a2}}{\partial \eta_{12}} &= \sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i2} = \sum_i \sum_j w_{ij} v_i A_{i2} Z_{i1}, \\
\frac{\partial U_{a3}}{\partial \eta_{12}} &= \sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i2} = \sum_i \sum_j w_{ij} v_i A_{i2} Z_{i2}, \\
\frac{\partial U_{b1}}{\partial \eta_{12}} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2}, \\
\frac{\partial U_{b2}}{\partial \eta_{12}} &= 0, \\
\frac{\partial U_{b3}}{\partial \eta_{12}} &= -\sum_i \sum_j w_{ij} A_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2}, \\
\frac{\partial U_{b4}}{\partial \eta_{12}} &= -\sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i1}, \\
\frac{\partial U_{b5}}{\partial \eta_{12}} &= -\sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i2}, \\
\frac{\partial U_{b6}}{\partial \eta_{12}} &= 0, \\
\frac{\partial U_{b7}}{\partial \eta_{12}} &= 0, \\
\frac{\partial U_{b8}}{\partial \eta_{12}} &= -\sum_i \sum_j w_{ij} A_{i2} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i1}, \\
\frac{\partial U_{b9}}{\partial \eta_{12}} &= -\sum_i \sum_j w_{ij} A_{i2} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i2}.
\end{aligned}
$$

For η_{21},

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \eta_{21}} &= \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} Z_{i1} = \sum_i \sum_j w_{ij} v_i Z_{i1}, \\
\frac{\partial U_{a2}}{\partial \eta_{21}} &= \sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} Z_{i1} = \sum_i \sum_j w_{ij} v_i Z_{i1}, \\
\frac{\partial U_{a3}}{\partial \eta_{21}} &= 0, \\
\frac{\partial U_{b1}}{\partial \eta_{21}} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i1} = -\sum_i \sum_j w_{ij} m_i Z_{i1}, \\
\frac{\partial U_{b2}}{\partial \eta_{21}} &= -\sum_i \sum_j w_{ij} A_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i1}, \\
\frac{\partial U_{b3}}{\partial \eta_{21}} &= -\sum_i \sum_j w_{ij} A_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i1}, \\
\frac{\partial U_{b4}}{\partial \eta_{21}} &= -\sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i1} = -\sum_i \sum_j w_{ij} m_i Z_{i1}, \\
\frac{\partial U_{b5}}{\partial \eta_{21}} &= 0, \\
\frac{\partial U_{b6}}{\partial \eta_{21}} &= -\sum_i \sum_j w_{ij} A_{i1} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i1}, \\
\frac{\partial U_{b7}}{\partial \eta_{21}} &= 0, \\
\frac{\partial U_{b8}}{\partial \eta_{21}} &= -\sum_i \sum_j w_{ij} A_{i2} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i1}, \\
\frac{\partial U_{b9}}{\partial \eta_{21}} &= 0.
\end{aligned}
$$

For η_{22},

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \eta_{22}} &= \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} Z_{i2} = \sum_i \sum_j w_{ij} v_i Z_{i2}, \\
\frac{\partial U_{a2}}{\partial \eta_{22}} &= 0, \\
\frac{\partial U_{a3}}{\partial \eta_{22}} &= \sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} Z_{i2} = \sum_i \sum_j w_{ij} v_i Z_{i2}, \\
\frac{\partial U_{b1}}{\partial \eta_{22}} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i2} = -\sum_i \sum_j w_{ij} m_i Z_{i2}, \\
\frac{\partial U_{b2}}{\partial \eta_{22}} &= -\sum_i \sum_j w_{ij} A_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i2}, \\
\frac{\partial U_{b3}}{\partial \eta_{22}} &= -\sum_i \sum_j w_{ij} A_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i2}, \\
\frac{\partial U_{b4}}{\partial \eta_{22}} &= 0, \\
\frac{\partial U_{b5}}{\partial \eta_{22}} &= -\sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i2} = -\sum_i \sum_j w_{ij} m_i Z_{i2}, \\
\frac{\partial U_{b6}}{\partial \eta_{22}} &= 0, \\
\frac{\partial U_{b7}}{\partial \eta_{22}} &= -\sum_i \sum_j w_{ij} A_{i1} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i2}, \\
\frac{\partial U_{b8}}{\partial \eta_{22}} &= 0, \\
\frac{\partial U_{b9}}{\partial \eta_{22}} &= -\sum_i \sum_j w_{ij} A_{i2} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i2}.
\end{aligned}
$$

For η_{31},

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \eta_{31}} &= \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i1} Z_{i1} = \sum_i \sum_j w_{ij} v_i A_{i1} Z_{i1}, \\
\frac{\partial U_{a2}}{\partial \eta_{31}} &= \sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i1} Z_{i1} = \sum_i \sum_j w_{ij} v_i A_{i1} Z_{i1}, \\
\frac{\partial U_{a3}}{\partial \eta_{31}} &= 0, \\
\frac{\partial U_{b1}}{\partial \eta_{31}} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i1}, \\
\frac{\partial U_{b2}}{\partial \eta_{31}} &= -\sum_i \sum_j w_{ij} A_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i1}, \\
\frac{\partial U_{b3}}{\partial \eta_{31}} &= 0, \\
\frac{\partial U_{b4}}{\partial \eta_{31}} &= -\sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i1}, \\
\frac{\partial U_{b5}}{\partial \eta_{31}} &= 0, \\
\frac{\partial U_{b6}}{\partial \eta_{31}} &= -\sum_i \sum_j w_{ij} A_{i1} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i1}, \\
\frac{\partial U_{b7}}{\partial \eta_{31}} &= 0, \\
\frac{\partial U_{b8}}{\partial \eta_{31}} &= 0, \\
\frac{\partial U_{b9}}{\partial \eta_{31}} &= 0.
\end{aligned}
$$

For η_{32},

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \eta_{32}} &= \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i1} Z_{i2} = \sum_i \sum_j w_{ij} v_i A_{i1} Z_{i2}, \\
\frac{\partial U_{a2}}{\partial \eta_{32}} &= 0, \\
\frac{\partial U_{a3}}{\partial \eta_{32}} &= \sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i1} Z_{i2} = \sum_i \sum_j w_{ij} v_i A_{i1} Z_{i2}, \\
\frac{\partial U_{b1}}{\partial \eta_{32}} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i2}, \\
\frac{\partial U_{b2}}{\partial \eta_{32}} &= -\sum_i \sum_j w_{ij} A_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i2}, \\
\frac{\partial U_{b3}}{\partial \eta_{32}} &= 0, \\
\frac{\partial U_{b4}}{\partial \eta_{32}} &= 0, \\
\frac{\partial U_{b5}}{\partial \eta_{32}} &= -\sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i2}, \\
\frac{\partial U_{b6}}{\partial \eta_{32}} &= 0, \\
\frac{\partial U_{b7}}{\partial \eta_{32}} &= -\sum_i \sum_j w_{ij} A_{i1} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i1} Z_{i2}, \\
\frac{\partial U_{b8}}{\partial \eta_{32}} &= 0, \\
\frac{\partial U_{b9}}{\partial \eta_{32}} &= 0.
\end{aligned}
$$

For η_{33},

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \eta_{33}} &= \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i2} Z_{i1} = \sum_i \sum_j w_{ij} v_i A_{i2} Z_{i1}, \\
\frac{\partial U_{a2}}{\partial \eta_{33}} &= \sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i2} Z_{i1} = \sum_i \sum_j w_{ij} v_i A_{i2} Z_{i1}, \\
\frac{\partial U_{a3}}{\partial \eta_{33}} &= 0, \\
\frac{\partial U_{b1}}{\partial \eta_{33}} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i1}, \\
\frac{\partial U_{b2}}{\partial \eta_{33}} &= 0, \\
\frac{\partial U_{b3}}{\partial \eta_{33}} &= -\sum_i \sum_j w_{ij} A_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i1}, \\
\frac{\partial U_{b4}}{\partial \eta_{33}} &= -\sum_i \sum_j w_{ij} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i1}, \\
\frac{\partial U_{b5}}{\partial \eta_{33}} &= 0, \\
\frac{\partial U_{b6}}{\partial \eta_{33}} &= 0, \\
\frac{\partial U_{b7}}{\partial \eta_{33}} &= 0, \\
\frac{\partial U_{b8}}{\partial \eta_{33}} &= -\sum_i \sum_j w_{ij} A_{i2} Z_{i1} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} Z_{i1} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i1}, \\
\frac{\partial U_{b9}}{\partial \eta_{33}} &= 0.
\end{aligned}
$$

For η_{34},

$$
\begin{aligned}
\frac{\partial U_{a1}}{\partial \eta_{34}} &= \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i2} Z_{i2} = \sum_i \sum_j w_{ij} v_i A_{i2} Z_{i2}, \\
\frac{\partial U_{a2}}{\partial \eta_{34}} &= 0, \\
\frac{\partial U_{a3}}{\partial \eta_{34}} &= \sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)}{[1+\exp(l_i - A_{i1}\xi_1 - A_{i2}\xi_2)]^2} A_{i2} Z_{i2} = \sum_i \sum_j w_{ij} v_i A_{i2} Z_{i2}, \\
\frac{\partial U_{b1}}{\partial \eta_{34}} &= -\sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i2}, \\
\frac{\partial U_{b2}}{\partial \eta_{34}} &= 0, \\
\frac{\partial U_{b3}}{\partial \eta_{34}} &= -\sum_i \sum_j w_{ij} A_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i2}, \\
\frac{\partial U_{b4}}{\partial \eta_{34}} &= 0, \\
\frac{\partial U_{b5}}{\partial \eta_{34}} &= -\sum_i \sum_j w_{ij} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i2}, \\
\frac{\partial U_{b6}}{\partial \eta_{34}} &= 0, \\
\frac{\partial U_{b7}}{\partial \eta_{34}} &= 0, \\
\frac{\partial U_{b1}}{\partial \eta_{34}} &= 0, \\
\frac{\partial U_{b1}}{\partial \eta_{34}} &= -\sum_i \sum_j w_{ij} A_{i2} Z_{i2} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} Z_{i2} = -\sum_i \sum_j w_{ij} m_i A_{i2} Z_{i2},
\end{aligned}
$$

Let $\mathbf{U}(\xi_0, \xi, \eta) = \begin{pmatrix} \mathrm{U}_{a1} \\ \mathrm{U}_{a2} \\ \mathrm{U}_{a3} \\ \mathrm{U}_{b1} \\ \vdots \\ \mathrm{U}_{a9} \end{pmatrix}$ and the gradient of $\mathbf{U}(\xi_0, \xi, \eta)$ is

$$\nabla \mathbf{U}(\xi_0, \xi, \eta) = \begin{pmatrix} \frac{\partial U_{a1}}{\partial \xi_0} & \dots & \frac{\partial U_{a1}}{\partial \eta_{34}} \\ \vdots & & \vdots \\ \frac{\partial U_{b9}}{\partial \xi_0} & \dots & \frac{\partial U_{b9}}{\partial \eta_{34}} \end{pmatrix}.$$

Then the variance is

$$Var(\xi_0, \hat{\xi}, \hat{\eta}) = [\nabla \mathbf{U}(\xi_0, \hat{\xi}, \hat{\eta})]^{-1} G [\nabla \mathbf{U}(\xi_0, \hat{\xi}, \hat{\eta})^T]^{-1},$$

where

$$G = \sum_{p=1}^{P} \frac{C_p}{C_p - 1} \sum_{i=1}^{C_p} (\mathbf{U}_{pi\cdot}(\xi_0, \hat{\xi}, \hat{\eta}) - \bar{\mathbf{U}}_{p\cdot\cdot}(\xi_0, \hat{\xi}, \hat{\eta}))(\mathbf{U}_{pi\cdot}(\xi_0, \hat{\xi}, \hat{\eta}) - \bar{\mathbf{U}}_{p\cdot\cdot}(\xi_0, \hat{\xi}, \hat{\eta}))^T.$$

In the SWASH study, p indexes geometric stratum (Rachuonyo District and Suba District) and i indexes clusters (schools within the two districts), $\mathbf{U}_{pi\cdot}(\xi_0, \hat{\xi}, \hat{\eta})$ is the sum of $\{\mathbf{U}_{pij}(\xi_0, \hat{\xi}, \hat{\eta})\}$ over j, and $\mathbf{U}_{p\cdot\cdot}(\xi_0, \hat{\xi}, \hat{\eta})$ is the sum of $\{\mathbf{U}_{pi\cdot}(\xi_0, \hat{\xi}, \hat{\eta})\}$ over i.

Next we consider derive the estimators (U_1 and U_2 respectively) of $E(Y_{ij}(a)|A_i = a)$ and $E(Y_{ij}(0)|A_i = a)$.

$$\begin{aligned} E(Y_{ij}(a)|A_i = a) &= E_{Z_i|A_i}(\mu(A_i, Z_i; \mu)) \\ &= E_{Z_i|A_i} expit(l_i), \end{aligned}$$

and

$$U_1 \stackrel{def}{=} \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} expit(l_i) I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}}.$$

$$
\begin{aligned}
E(Y_{ij}(0)|A_i = a) &= E_{Z_i|A_i}[h^{-1}(h(expit(l_i)) - A_i\xi)] \\
&= E_{Z_i|A_i}[expit(l_i - A_i\xi)],
\end{aligned}
$$

and

$$
U_2 \stackrel{def}{=} \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} expit(l_i - A_i\xi) I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}}.
$$

Delta Method: If a consistent estimator T converges in probability to its true value τ, then after applying central limit theorem

$$
\sqrt{n}(T - \tau) \xrightarrow{D} N(0, \Sigma),
$$

where n is number of observations and Σ is the covariance. Any function h which $\nabla h(\tau)$ exists and is non-zero valued, then

$$
\sqrt{n}(h(T) - h(\tau)) \xrightarrow{D} N(0, \nabla h(\tau)^T \Sigma \nabla h(\tau)).
$$

In our case, let

$$
\tau = \begin{pmatrix} E(Y_{ij}(a)|A_i = a) \\ E(Y_{ij}(0)|A_i = a) \end{pmatrix},
$$

and

$$
T = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix},
$$

as the respect estimator of τ. We obtain the following partial derivatives,

$$
\begin{aligned}
\frac{\partial U_1}{\partial \xi_0} &= 0 \\
\frac{\partial U_1}{\partial \xi_1} &= 0 \\
\frac{\partial U_1}{\partial \xi_2} &= 0 \\
\frac{\partial U_1}{\partial \eta_{10}} &= \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_1}{\partial \eta_{11}} &= \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_1}{\partial \eta_{12}} &= \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_1}{\partial \eta_{21}} &= \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i1} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_1}{\partial \eta_{22}} &= \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} Z_{i2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_1}{\partial \eta_{31}} &= \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} Z_{i1} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_1}{\partial \eta_{32}} &= \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i1} Z_{i2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_1}{\partial \eta_{33}} &= \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} Z_{i1} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_1}{\partial \eta_{34}} &= \frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i)}{[1+\exp(l_i)]^2} A_{i2} Z_{i2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial U_2}{\partial \xi_0} &= 0 \\
\frac{\partial U_2}{\partial \xi_1} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} A_{i1} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_2}{\partial \xi_3} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} A_{i2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_2}{\partial \eta_{10}} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_2}{\partial \eta_{11}} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} A_{i1} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_2}{\partial \eta_{12}} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} A_{i2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_2}{\partial \eta_{21}} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} Z_{i1} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_2}{\partial \eta_{22}} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} Z_{i2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_2}{\partial \eta_{31}} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} A_{i1} Z_{i1} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_2}{\partial \eta_{32}} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} A_{i1} Z_{i2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_2}{\partial \eta_{33}} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} A_{i2} Z_{i1} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}} \\
\frac{\partial U_2}{\partial \eta_{34}} &= -\frac{\sum_{\text{all possible z}} \sum_i \sum_j w_{ij} \frac{\exp(l_i - A_i\xi)}{[1+\exp(l_i - A_i\xi)]^2} A_{i2} Z_{i2} I\{A_i = a, Z_i = z\}}{\sum_i \sum_j w_{ij} I\{A_i = a\}}
\end{aligned}
$$

Therefore,

$$\nabla h(T) = \begin{pmatrix} \frac{\partial U_1}{\partial \xi_0} & \frac{\partial U_1}{\partial \xi_1} & \frac{\partial U_1}{\partial \xi_2} & \frac{\partial U_1}{\partial \eta_{10}} & \frac{\partial U_1}{\partial \eta_{11}} & \frac{\partial U_1}{\partial \eta_{12}} & \frac{\partial U_1}{\partial \eta_{21}} & \frac{\partial U_1}{\partial \eta_{22}} & \frac{\partial U_1}{\partial \eta_{31}} & \frac{\partial U_1}{\partial \eta_{32}} \\ \frac{\partial U_1}{\partial \eta_{33}} & \frac{\partial U_1}{\partial \eta_{34}} & & & & & & & & \\ \frac{\partial U_2}{\partial \xi_0} & \frac{\partial U_2}{\partial \xi_1} & \frac{\partial U_2}{\partial \xi_2} & \frac{\partial U_2}{\partial \eta_{10}} & \frac{\partial U_2}{\partial \eta_{11}} & \frac{\partial U_2}{\partial \eta_{12}} & \frac{\partial U_2}{\partial \eta_{21}} & \frac{\partial U_2}{\partial \eta_{22}} & \frac{\partial U_2}{\partial \eta_{31}} & \frac{\partial U_2}{\partial \eta_{32}} \\ \frac{\partial U_2}{\partial \eta_{33}} & \frac{\partial U_2}{\partial \eta_{34}}. & & & & & & & & \end{pmatrix}.$$

Then using delta method,

$$Var(T) = \nabla h(T) Var(\hat{\xi}_0, \hat{\xi}, \hat{\eta}) \nabla h(T)^T.$$

Now we turn to focus on RR and RD. Let

$$h^* \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} U_1/U_2 \\ U_1 - U_2 \end{pmatrix},$$

then the gradient is

$$\nabla h^* \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{U_2} & -\frac{U_1}{U_2} \\ 1 & -1 \end{pmatrix}.$$

Use delta method again, we can obtain the variance of RR and RD

$$Var \begin{pmatrix} RR \\ RD \end{pmatrix} = \nabla h^* \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}^T Var(T) \nabla h^* \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}.$$

REFERENCES

[1] Agresti, A. *Categorical data analysis*, vol. 359. John Wiley and Sons, 2002.

[2] Andersen, E.B. *Discrete statistical models with social science applications*. North-Holland Amsterdam, 1980.

[3] Angrist, J. and Krueger, A.B. "Instrumental variables and the search for identification: From supply and demand to natural experiments." Tech. rep., National Bureau of Economic Research, 2001.

[4] Angrist, J.D., Imbens, G.W., and Rubin, D.B. "Identification of causal effects using instrumental variables." *Journal of the American Statistical Association* 91 (1996).434: 444–455.

[5] Bao, Y., Duan, N., and Fox, S.A. "Is some provider advice on smoking cessation better than no advice? An instrumental variable analysis of the 2001 National Health Interview Survey." *Health services research* 41 (2006).6: 2114–2135.

[6] Beck, C.A., Penrod, J., Gyorkos, T.W., Shapiro, S., and Pilote, L. "Does aggressive care following acute myocardial infarction reduce mortality? Analysis with instrumental variables to compare effectiveness in Canadian and United States patient populations." *Health Services Research* 38 (2003).6 Pt 1: 1423–1440.

[7] Begg, M.D. and Parides, M.K. "Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data." *Statistics in Medicine* 22 (2003).16: 2591–2602.

[8] Berlin, J.A., Kimmel, S.E., Have, T.R.T., and Sammel, M.D. "An empirical comparison of several clustered data approaches under confounding due to cluster effects in the analysis of complications of coronary angioplasty." *Biometrics* 55 (1999).2: 470–476.

[9] Binder, D.A. "On the variances of asymptotically normal estimators from complex surveys." *International Statistical Review/Revue Internationale de Statistique* 51 (1983).3: 279–292.

[10] Botman, S.L., Moore, T.F., Moriarity, C.L., and V.L., Parsons. "Design and estimation for the National Health Interview Survey, 1995–2004." *Center for Health Statistics, Vital and Health Statistics* 2 (2000).130: 1–31.

[11] Bound, J., Jaeger, D.A., and Baker, R.M. "Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak." *Journal of the American statistical association* (1995): 443–450.

[12] Breslow, N.E., Day, N.E., Halvorsen, K.T., Prentice, R.L., and Sabai, C. "Estimation of multiple relative risk functions in matched case-control studies." *American Journal of Epidemiology* 108 (1978).4: 299–307.

[13] Brooks, J.M., Chrischilles, E.A., Scott, S.D., and Chen-Hardee, S.S. "Was breast conserving surgery underutilized for early stage breast cancer? Instrumental variables evidence for stage II patients from Iowa." *Health services research* 38 (2003).6p1: 1385–1402.

[14] Brumback, B.A., Dailey, A.B., He, Z., Brumback, L.C., and Livingston, M.D. "Efforts to adjust for confounding by neighborhood using complex survey data." *Statistics in medicine* 29 (2010).18: 1890–1899.

[15] Brumback, B.A. and He, Z. "Adjusting for confounding by neighborhood using complex survey data." *Statistics in Medicine* 30 (2011).9: 965–972.

[16] Bun, M.J.G. and Windmeijer, F. "A comparison of bias approximations for the two-stage least squares (2SLS) estimator." *Economics Letters* 113 (2011).1: 76–79.

[17] Casella, G. and Berger, R.L. *Statistical inference*. Duxbury Press, 2001.

[18] Glymour, M.M., Tchetgen, E.J.T., and Robins, J.M. "Credible mendelian randomization studies: approaches for evaluating the instrumental variable assumptions." *American journal of epidemiology* 175 (2012).4: 332–339.

[19] Goetgeluk, S. and Vansteelandt, S. "Conditional generalized estimating equations for the analysis of clustered and longitudinal data." *Biometrics* 64 (2008).3: 772–780.

[20] Graubard, B.I. and Korn, E.L. "Conditional logistic regression with survey data." *Statistics in Biopharmaceutical Research* 3 (2011).2: 398–408.

[21] Greenland, S. "An introduction to instrumental variables for epidemiologists." *International Journal of Epidemiology* 29 (2000).4: 722–729.

[22] ———. "A review of multilevel theory for ecologic analyses." *Statistics in Medicine* 21 (2002).3: 389–395.

[23] Greenland, S. and Morgenstern, H. "Confounding in health research." *Annual Review of Public Health* 22 (2001).1: 189–212.

[24] Greenland, S. and Robins, J.M. "Identifiability, exchangeability, and epidemiological confounding." *International Journal of Epidemiology* 15 (1986).3: 413–419.

[25] Greenland, S., Robins, J.M., and Pearl, J. "Confounding and collapsibility in causal inference." *Statistical Science* (1999): 29–46.

[26] Grilli, L. and Pratesi, M. "Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs." *Survey methodology* 30 (2004).1: 93–104.

[27] Grilli, L. and Rampichini, C. "Measurement error in multilevel models with sample cluster means." Tech. rep., Electronic Working Papers 6/2009, Department of Statistics-University of Florence, 2009.

[28] He, Z. and Brumback, B.A. "An Equivalence of Conditional and Unconditional Maximum Likelihood Estimators via In." *Communications in Statistics – Theory and Methods* (in press).

[29] Hernán, M.Á., Brumback, B., and Robins, J.M. "Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men." *Epidemiology* 11 (2000).5: 561–570.

[30] Hernán, M.A. and Robins, J.M. "Instruments for causal inference: An epidemiologist's dream?" *Epidemiology* (2006): 360–372.

[31] Holt, D. and Smith, TMF. "Post stratification." *Journal of the Royal Statistical Society. Series A (General)* (1979): 33–46.

[32] Korn, E.L. and Graubard, B.I. *Analysis of health surveys*. Wiley: New York, 1999.

[33] Lee, E.S. and Forthofer, R.N. *Analyzing complex survey data*. Sage Publications, Inc, 2006.

[34] Lehtonen, R. and Pahkinen, E. *Practical methods for design and analysis of complex surveys*, vol. 10. Wiley, 2004.

[35] Liang, K.Y. "Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models." *Biometrics* (1987): 289–299.

[36] Liang, K.Y. and Zeger, S.L. "Longitudinal data analysis using generalized linear models." *Biometrika* 73 (1986).1: 13–22.

[37] Lindsay, B., Clogg, C.C., and Grego, J. "Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis." *Journal of the American Statistical Association* (1991): 96–107.

[38] Little, R.J., Long, Q., and Lin, X. "A comparison of methods for estimating the causal effect of a treatment in randomized clinical trials subject to noncompliance." *Biometrics* 65 (2009).2: 640–649.

[39] Little, R.J.A. "Post-stratification: A modeler's perspective." *Journal of the American Statistical Association* (1993): 1001–1012.

[40] Localio, A.R., Berlin, J.A., and Have, T.R.T. "Confounding due to cluster in multicenter studies – causes and cures." *Health Services and Outcomes Research Methodology* 3 (2002).3: 195–210.

[41] Longford, NT. “Model-based variance estimation in surveys with stratified clustered design.” *Australian Journal of Statistics* 38 (1996).3: 333–352.

[42] Martens, E.P., Pestman, W.R., De Boer, A., Belitser, S.V., and Klungel, O.H. “Instrumental variables: application and limitations.” *Epidemiology* 17 (2006).3: 260.

[43] National Center for Health Statistics. “National Health Interview Survey (NHIS): 2005 data release (online).” http://www.cdc.gov/nchs/nhis/nhis_2005_data_release.htm, last accessed on Sep. 18, 2011.

[44] Neuhaus, J.M. and Kalbfleisch, J.D. “Between-and within-cluster covariate effects in the analysis of clustered data.” *Biometrics* (1998): 638–645.

[45] Neuhaus, J.M., Kalbfleisch, J.D., and Hauck, W.W. “Conditions for consistent estimation in mixed-effects models for binary matched-pairs data.” *Canadian Journal of Statistics* 22 (1994).1: 139–148.

[46] Neuhaus, J.M. and McCulloch, C.E. “Separating between-and within-cluster covariate effects by using conditional and partitioning methods.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (2006).5: 859–872.

[47] Neyman, J. and Scott, E.L. “Consistent estimates based on partially consistent observations.” *Econometrica: Journal of the Econometric Society* (1948): 1–32.

[48] Pearl, J. *Causality: models, reasoning and inference*. Cambridge Univ Press, 2000.

[49] Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash, J. “Weighting for unequal selection probabilities in multilevel models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60 (1998).1: 23–40.

[50] Pilote, L., Beck, C.A., Eisenberg, M.J., Humphries, K., Joseph, L., Penrod, J.R., and Tu, J.V. “Comparing invasive and noninvasive management strategies for acute myocardial infarction using administrative databases.” *American Heart Journal* 155 (2008).1: 42–48.

[51] Rabe-Hesketh, S. and Skrondal, A. “Multilevel modelling of complex survey data.” *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169 (2006).4: 805–827.

[52] Rasch, G. *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press (Chicago), 1980.

[53] Remington, P.L., Smith, M.Y., Williamson, D.F., Anda, R.F., Gentry, E.M., and Hogelin, G.C. “Design, characteristics, and usefulness of state-based behavioral risk factor surveillance: 1981-87.” *Public Health Reports* 103 (1988).4: 366–375.

[54] Rice, K. “Equivalence between conditional and random-effects likelihoods for pair-matched case-control studies.” *Journal of the American Statistical Association* 103 (2008).481: 385–396.

[55] Rice, K.M. “Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications.” *Journal of the American Statistical Association* 99 (2004).466: 510–522.

[56] Robins, J.M. “Correcting for non-compliance in randomized trials using structural nested mean models.” *Communications in Statistics-Theory and methods* 23 (1994).8: 2379–2412.

[57] Severini, T.A. “On the relationship between Bayesian and non-Bayesian elimination of nuisance parameters.” *Statistica Sinica* 9 (1999): 713–724.

[58] Skinner, C.J., Holt, D., and Smith, T.M.F. *Analysis of complex surveys*. John Wiley & Sons, 1989.

[59] Stock, J.H. and Trebbi, F. “Retrospectives: Who invented instrumental variable regression?” *The Journal of Economic Perspectives* 17 (2003).3: 177–194.

[60] Stukel, T.A., Fisher, E.S., Wennberg, D.E., Alter, D.A., Gottlieb, D.J., and Vermeulen, M.J. “Analysis of observational studies in the presence of treatment selection bias.” *The journal of the American Medical Association* 297 (2007).3: 278–285.

[61] Ten Have, T.R., Normand, S.L.T., Marcus, S.M., Brown, C.H., Lavori, P., and Duan, N. “Intent-to-treat vs. non-intent-to-treat analyses under treatment non-adherence in mental health randomized trials.” *Psychiatric annals* 38 (2008).12: 772–783.

[62] Vansteelandt, S. and Goetghebeur, E. “Causal inference with generalized structural mean models.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65 (2003).4: 817–835.

[63] Verbeke, G., Spiessens, B., and Lesaffre, E. “Conditional linear mixed models.” *The American Statistician* 55 (2001).1: 25–34.

[64] von Hinke Kessler Scholder, S., Propper, C., Lawlor, D., Windmeijer, F., and Davey Smith, G. “Genetic markers as instrumental variables: an application to child fat mass and academic achievement.” (2010).

[65] Wehby, G.L., Ohsfeldt, R.L., and Murray, J.C. “Mendelian randomizationequals instrumental variable analysis with genetic instruments.” *Statistics in medicine* 27 (2008).15: 2745–2749.

[66] Wright, P.G. *Tariff on animal and vegetable oils*. 26. Macmillan Company, New York, 1928.

BIOGRAPHICAL SKETCH

Zhulin He was born in Tianjin, China. After graduating from Tianjin Nankai High School in 2002, she entered Nankai University for her four year undergraduate study at the School of Mathematical Science, major in Statistics. From 2006-2008, she attended University of Calgary in Canada. She finished her thesis on “Estimation for Censored Single-Index Models Using MAWVE and OPWG Methods” under the guidance of her advisor, Dr. Xuewen Lu, and received the degree of Master of Science in Statistics in 2008. During the following years, she studied biostatistics under her advisor, Dr. Babette Brumback’s help. She mainly studied and worked on causal inference with complex sampling designs. In the summer of 2012, she received her Ph.D. from the University of Florida.