Estimating Equations Approaches to Nuisance Parameters and Outcome-Dependent Sampling Problems for Marginal Regression Models and Generalized Linear Mixed Models When Outcomes Are Correlated

Kunthel By

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

Chapel Hill
2011

Approved by:

Advisor: Dr. Bahjat F. Qaqish

Reader: Dr. Lloyd J. Edwards

Reader: Dr. Robert C. Millikan

Reader: Dr. John S. Preisser

Reader: Dr. Mark A. Weaver

Abstract

KUNTHEL BY: Estimating Equations Approaches to Nuisance Parameters and Outcome-Dependent Sampling Problems for Marginal Regression Models and Generalized Linear Mixed Models When Outcomes Are Correlated. (Under the direction of Dr. Bahjat F. Qaqish.)

For marginal regression models having cluster-specific intercepts, the number of model parameters grows with the sample size so that GEE is not feasible. A solution is to impose a mixing distribution on the intercepts which leads to generalized linear mixed models (GLMMs) whose parameters have different interpretations than marginal models. When GLMM assumptions are not met, parameter estimates are generally biased. A simple procedure for constructing estimating equations is proposed that enables consistent estimation of parameters associated with cluster-varying covariates and is applicable regardless of whether the cluster-specific intercept is treated as fixed or random. The proposed procedure is shown to work for the identity and log links but not for the logit link. Connections to conditional likelihoods, the Cox model, projected score, and adjusted profile likelihoods are discussed. It is shown that our estimating equations can be implemented with minimal programming effort using existing software. We show that a connection exists between biased sampling based on cluster totals and regression models with cluster-specific intercepts. This connection leads naturally to our estimation procedure. Regression parameters associated with cluster-varying covariates can be consistently estimated using our estimating function even when sampling rates are unknown. An estimation procedure based on the double-pair design and an estimating function for a 1-1 matching design are shown to be special cases of our procedure. Risk ratio estimation is possible for case-control studies when family members are chosen as controls.

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Bahjat Qaqish, for his guidance throughout this project. From inception to completion, he was instrumental in developing my understanding of the subject matter. I can no longer look at a data set without asking "how was the data collected" and what was the "design"? His insistence on precision, brevity, clarity, and simplicity has force me to think hard before I verbalize and pen my thoughts (at times, I ignored his advice and plunge headlong into idiocy). I owe him a great deal of gratitude. As a testament to his kindness and belief in people, he accepted me as his student (perhaps grudgingly - who knows) despite a moronic performance (read ZERO!) on an exam when I first took his class.

I am also appreciative of my committee members. Their responses to both my verbal and written communications have resulted in a manuscript that not only is more polished but also clearer and (hopefully) more coherent. Bob Millikan kindly provided data from the Carolina Breast Cancer Study which enabled us to broaden the application of our results.

Communicating technical ideas has always been one of my weaknesses. Through the guidance and insistence of the following individuals, I have gotten substantially better: Drs. Qaqish, Preisser, Edwards, and Michael Harvey.

I cannot stress enough how difficult my circumstances would be had it not been for the financial support (through GRAs, Training Grants, and outright generosity) of the following individuals: Drs. Amy Herring, Lloyd Edwards, Bahjat Qaqish, John Preisser, Gary Slade, Ed Davis, and Ms. Anne Jacob.

Also, I would like to thank the bios staff. Their excellent work has enabled me to focus on my work rather than worry about administrative stuff.

To Melissa and Herb Ubbens, Vonn and Natalie Walter, you guys are like family. I am especially grateful to you for dinner company, game nights, and for taking care of Kendall when I am scrambling to meet deadlines. I am also thankful to Tania Osborne whose sense of humor and sympathy helped me out through those dark days. To all, you have contributed in some special way to the success of this project. I can't think of better company to complain about school (and everything else too).

To Kendall, your shenanigans and sense of humor have brought me great joy. And Flora, thank you for being such a great mother to my son, for taking over the responsibilities of running the household, for tolerating my absence, and for repeatedly reading my manuscript for grammatical and spelling errors. Your unwavering love and support kept me on course.

Finally, without the sacrifices of my mother and my late father, it would not have been possible to be where I am today.

Table of Contents

List Of Abbreviations . . . . . **xi**

1 Introduction **1**

1.1 Nuisance Parameters 1

1.2 Biased Sampling 2

2 Nuisance Parameters **3**

2.1 Introduction 3

2.2 Notations & Conventions 6

2.3 Motivation 7

2.3.1 Meta-Analysis 7

2.3.2 Multi-center Studies 8

2.3.3 Stratified Sampling Designs 8

2.3.4 Repeated Measures/Longitudinal Studies (RML) 8

2.4 Previous Works 9

2.4.1 Conditional Likelihood 9

2.4.2 Conditional GEE (CGEE) 9

2.4.3 Optimally-Weighted Estimating Functions 12

2.4.4 Projected Estimating Function Method 13

2.4.5 Cox-Reid Adjusted Profile Estimating Function Method 15

2.4.6 Barndorf-Nielsen Profile Estimating Function Method 16

2.4.7 Summary 18

2.5 Estimating Functions in the Presence of Cluster-Specific Nuisance Parameters: A Motivation 20

2.6 Identity Link 22

2.6.1 Asymptotic Properties of $\widehat{\boldsymbol{\beta}}$ 23

2.6.2 Estimation 25

2.6.3 Comments on Efficiency 26

2.6.4 Situations Where This Approach is Appropriate 26

2.7 Log Link 26

2.7.1 Asymptotic Properties of $\widehat{\boldsymbol{\beta}}$ . . . 28
2.7.2 Other Variance Functions . . . 28
2.8 A Generalization . . . 29
2.8.1 Log Link and Gamma Variance Function . . . 30
2.8.2 Log Link and Inverse Gaussian Variance Function . . . 31
2.9 Connections To the Cox Model . . . 32
2.10 Connections to the Conditional Likelihood . . . 35
2.10.1 Linear Regression . . . 35
2.10.2 Poisson Regression . . . 35
2.11 Connections to De-Sensitization Methods . . . 36
2.11.1 Projected Estimating Function . . . 36
2.11.2 Cox-Reid Type Adjusted Profile Estimating Function . . . 37
2.11.3 The Binomial Variance Function . . . 37
2.12 Conclusion . . . 39
3 Biased Sampling, Clustered Binary Data, and Regression Models . . . 41
3.1 Introduction . . . 41
3.2 Notations & Conventions . . . 42
3.3 Biased Sampling . . . 45
3.4 Previous Works . . . 47
3.4.1 Proband Design . . . 47
3.4.2 Stratified Sampling . . . 49
3.4.3 Sampling Based on the Total . . . 52
3.5 Sampling Based on the Total: A Closer Look . . . 56
3.6 Characterizing the Outcome-Dependent Sampling . . . 57
3.7 Marginal Models . . . 58
3.7.1 Sampling Clusters Based on the Number of Events . . . 60
3.7.2 Sampling Clusters Exhibiting Variation . . . 63
3.7.3 Summary of Outcome-Dependent Sampling For Marginal Models . . . 65
3.8 Connections to the Double-Pair Design . . . 65
3.9 Simulations: Marginal Models and Outcome-Dependent Sampling . . . 68
3.9.1 Log Link . . . 69
3.9.2 Logit Link . . . 72
3.10 Generalized Linear Mixed Models (GLMMs) . . . 73
3.10.1 Sampling Clusters Based on the Number of Events . . . 78
3.10.2 Sampling Clusters Exhibiting Variation . . . 79

3.11 Simulations: GLMMs and Outcome-Dependent Sampling 79
3.11.1 Log Link 80
3.11.2 Logit Link 82
3.12 Summary of Simulation Results 87
3.13 Discussion 88
4 Applications 93
4.1 Introduction 93
4.2 Birth Weight and Risk of Death 94
4.3 Helmet Effectiveness 97
4.4 Carolina Breast Cancer Study 99
5 Future Research 103
5.1 Introduction 103
5.2 Between-Within Decomposition 103
5.3 Comparisons With Other Methods 104
5.4 Biased Sampling and Logit-Normal GLMMs 105
A 106
A.1 Evans' Estimator 106
A.2 Twins Analyses 108
A.3 Conditional Likelihood 109
A.4 Proof Of Propositions 111
Bibliography 127

List of Tables

3.1 Double-Pair Design 66

3.2 Log link marginal model simulation study: estimation of biased sample data using the method of By and Qaqish 70

3.3 Log link marginal model simulation study: estimation based on all clusters using GEE 71

3.4 Log link marginal model simulation study: efficiency calculations 71

3.5 Log link marginal model simulation study: estimation of biased sample data using naive GEE 72

3.6 Logit link marginal model simulation study: estimation of biased data using the method of By and Qaqish 74

3.7 Logit link marginal model simulation study: estimation based on all clusters using GEE 74

3.8 Logit link marginal model simulation study: efficiency calculations 74

3.9 Logit link marginal model simulation study: estimation of biased sample data using naive GEE 75

3.10 Log link GLMM simulation study: estimation of biased data using the method of By and Qaqish 81

3.11 Log link GLMM simulation study: estimation of biased data using the method of between-within decomposition 81

3.12 Log link GLMM simulation study: estimation of biased data by naively fitting a GLMM 82

3.13 Logit link GLMM simulation study: estimation of model 1 parameters using the method of By and Qaqish 83

3.14 Logit link GLMM simulation study: estimation of model 1 parameters using the method of between-within decomposition 84

3.15 Logit link GLMM simulation study: estimation of model 1 parameters ignoring sampling scheme 85

3.16 Logit link GLMM simulation study: estimation of model 2 parameters using the method of By and Qaqish 86

3.17 Logit link GLMM simulation study: estimation of model 2 parameters using the method of between-within decomposition 86

3.18 Logit link GLMM simulation study: estimation of model 2 parameters ignoring the sampling scheme 87

3.19 Logit link GLMM simulation study: estimation of model 3 parameters using the method of By and Qaqish 91

3.20 Logit link GLMM simulation study: estimation of model 3 parameters ignoring the sampling scheme 92

4.1 Helmet effectiveness data 98

4.2 Comparison of estimates between Greenland's estimating function and the method of By and Qaqish 99

A.1 Twin pairs data from 1995 to 2000 matched multiple birth file 108

A.2 Sample of twin pairs data obtained from biased sampling scheme (4.2) 109

A.3 A comparison of estimates between three methods based on biased data obtained from sampling scheme (4.2) 109

A.4 Sample of twin pairs data obtained from biased sampling scheme (4.4) 109

A.5 A comparison of estimates between three methods based on biased data obtained from sampling scheme (4.4) 110

List Of Abbreviations

GEE	Generalized Estimating Equations
CGEE	Conditional GEE
GLMM	Generalized Linear Mixed Model
PEF	Profile Estimating Function
APrEF	Adjusted Profile Estimating Function
PNR	Probability dependent on the total Number of Responses
VP	Variable Probability
VPS	Variability Probability Sampling

Chapter 1

Introduction

1.1 Nuisance Parameters

Perhaps for as long as man invented statistical models, there exists certain model parameters that inherently of no interest to anyone and are known as nuisance parameters. These parameters exist to ensure model validity. The side-effects of nuisance parameters are two-fold (McCullagh and Nelder, 1989). First, as the number of nuisance parameters grows, no theory exists to guarantee the consistency of both the estimates of the parameter of interest and the nuisance parameters. Even if the estimates enjoy consistency, there is no guarantee that they will be efficient. Second, solving estimating functions or maximizing likelihoods with many parameters is numerically difficult. Many procedures have been suggested for addressing the nuisance parameter problem. These include data transformation, conditional likelihood, conditional score (Lindsay, 1982), projected score (Waterman and Lindsay, 1996; Small and McLeish, 1989), corrected profile likelihood (Cox and Reid, 1987; Barndorff-Nielsen, 1983), and estimating equations (Severini, 2002; Godambe, 1991).

Our interest in the nuisance parameter problem does not lie with independent data. Rather, we imagine a population of vectors of correlated outcomes and we wish to study the relationship between outcomes and covariates. Chapter 2 of this dissertation explores the nature of this relationship through regression models that contain cluster-specific nuisance parameters. After discussing existing methods, we will propose novel estimating equations for estimating certain parameters of interest.

1.2 Biased Sampling

Much of the existing statistical procedures were developed under the assumption of random sampling. In practice, this assumption rarely holds. Data that is not obtained via random sampling is said to have arisen from a biased sampling scheme. For example, case-control studies do not collect data that is representative of the population. Special methods or adjustments to existing methods are usually needed in the presence of biased sampling. For independent data, there has been extensive research on statistical procedures that adjust for the sampling scheme. It is not within the scope of this dissertation to add to this. Our interest lies with statistical models for correlated binary outcomes. We consider a particular type of sampling scheme based on cluster totals, a form of outcome-dependent sampling. Chapter 3 provides a formal description of outcome-dependent sampling based on cluster totals and examines the relationship between the model induced by the sampling scheme (the biased sampling model) and the specified population model. Connections are then made between the biased sampling model and the nuisance parameter problem described earlier.

Applications and Future Research

Chapter 4 considers applications of our results to three data sets while chapter 5 enumerates possibilities for future work on matters related to this dissertation topic.

Chapter 2

Regression Models, Cluster-Specific Nuisance Parameters, and Estimating Equations

2.1 Introduction

Consider the following class of generalized linear models:

$$h(\mu_{ij}) = \lambda_i + \beta_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{w}_i^\top \boldsymbol{\gamma}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, n_i \tag{2.1}$$

where h is the link function, i indexes clusters, and j indexes observations within clusters

- $\boldsymbol{x}_{ij}$ is a $p \times 1$ vector of covariates that vary within the cluster
- $\boldsymbol{w}_i$ is a $q \times 1$ vector of covariates that are cluster-constant
- λ_i is a cluster-specific intercept
- $\mu_{ij} = E[Y_{ij} | \lambda_i, \boldsymbol{X}_i, \boldsymbol{W}_i]$
- $\boldsymbol{X}_i$ is a $n_i \times p$ matrix of cluster-varying covariates
- $\boldsymbol{W}_i$ is a $n_i \times q$ matrix of cluster-constant covariates

Assume that the parameter of interest is $\boldsymbol{\beta}$. We can rewrite (2.1) as

$$h(\mu_{ij}) = \delta_i + \beta_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} \tag{2.2}$$

where $\delta_i = \lambda_i + \boldsymbol{w}_i^\top \boldsymbol{\gamma}$. Since only $\boldsymbol{\beta}$ is the parameter of interest, δ_i can be viewed as a nuisance parameter.

For marginal regression models, δ_i is treated as unknown constants and GEE may be used to estimate $\boldsymbol{\beta}$. However, since the number of nuisance parameters is increasing with the sample size, it is not clear that $\widehat{\boldsymbol{\beta}}$ will be consistent. This problem may be viewed as an extension of the Neyman-Scott problem (Neyman and Scott, 1948).

One way of overcoming this setback is to impose a distribution on δ_i; for example, $\delta_i \sim G(\boldsymbol{\theta})$. In practice, this is typically done by assuming that $\delta_i \sim N(\delta, \sigma_\delta^2)$. This puts us in the framework of generalized linear mixed models (GLMMs) and the interpretation of $\boldsymbol{\beta}$ is no longer the same as when δ_i is viewed as a constant. There are two potential shortcomings of this approach. First is the issue of robustness to misspecification of the mixing distribution. While the work of Neuhaus et al. (1992) suggests that the effect of misspecifying the mixing distribution is small, the works of Heagerty and Kurland (2001) and Heagerty and Zeger (2000) suggest otherwise. They showed that estimates of regression parameters can be biased when the mixing distribution is misspecified. For example, if the variance of the random effects depends on cluster-level covariates, $\widehat{\boldsymbol{\beta}}$ can be substantially biased. Second is the issue of robustness to violations of GLMM assumptions. Neuhaus and McCulloch (2006) discussed two examples where the random effects are correlated with the covariates. When this happens, parameter estimates from naively fitting GLMMs are generally inconsistent. To remedy this problem, they proposed the between-within covariate decomposition technique and showed by simulations that estimates of the slope parameters are unbiased. While the between-within decomposition method is simple to implement with existing software, a serious drawback is the interpretation of parameter estimates. Consider the following linear predictor associated with a GLMM:

$$\eta_{ij} = b_i + \beta_0 + \beta_1 x_{ij} \ .$$

Here, β_1 can be interpreted as a contrast: given b_i, β_1 is the change in the mean (in the scale of the link function) for a one unit increase in x_{ij}. In the between-within approach, the covariate

and the slope are decomposed as follows:

$$\eta_{ij} = b_i + \beta_0 + \beta_B \bar{x}_i + \beta_W (x_{ij} - \bar{x}_i) \tag{2.3}$$

where $\bar{x}_i$ is the simple average of the covariate for the i-th cluster: $(\bar{x}_i := n_i^{-1} \sum_{j=1}^{n_i} x_{ij})$. The parameter β_B provides some measure of between-cluster effect while the parameter β_W provides some measure of within-cluster effect. Through their simulation study, Neuhaus and McCulloch (2006) showed that $\hat{\beta}_W$ is unbiased for β_1 while $\hat{\beta}_B$ is biased. Thus, when the random effects are correlated with the covariates, they suggested fitting a GLMM using the between-within linear predictor and using $\hat{\beta}_W$ as an estimate of β_1. However, this approach raises two conceptual difficulties. First, does β_W have the same interpretation as β_1? Second, what is the interpretation of β_B? The paper does not clarify this issue. For fixed $\bar{x}_i$, β_W may be interpreted as the change in the mean (in the scale of the link function) for a unit increase in $(x_{ij} - \bar{x}_i)$. But a one unit increase in $(x_{ij} - \bar{x}_i)$ is impossible if $\bar{x}_i$ and all $\boldsymbol{x}_{ik},\ k \neq j$, are fixed. In addition, the between-within approach continues to be susceptible to bias when the mixing distribution is misspecified.

In summary, both the GLMM and GEE approaches to model (2.2) have shortcomings. The GLMM approach reduces the number of nuisance parameters from K to that of the dimension of the parameter vector that describes the mixing distribution but is subject to bias if the mixing distribution is misspecified or if the random effects are correlated with covariates. On the other hand, the GEE approach is not feasible since the number of nuisance parameters is increasing with K. Under this backdrop, we seek an alternative estimation procedure that avoids having to specify the mixing distribution in the GLMM setup and address the nuisance parameter problem in the GEE setup. In particular, this dissertation proposes new estimating functions that are simple to implement and free of nuisance parameters.

This chapter is organized as follows. Section 2.2 introduces the necessary notation and terminology. Section 2.3 provides some motivations for how model (2.1) may arise. A discussion of previous works related to the previously discussed problems is provided in section 2.4. Section 2.5 provides a motivation that leads to our estimating functions which are treated separately for

the identity (Section 2.6) and log (Section 2.7) links. We highlight some difficulties associated with our procedure and the logit link function. Section 2.9 discusses a connection between our estimating function and the Cox model. Parallels are drawn between the method of conditional likehood and our estimating functions in section 2.10. Connections between our estimating functions and the de-sensitized estimating functions of other authors are discussed in section 2.11 with a view towards showing that our procedure is much simpler. Finally, section 2.12 provides a summary of the salient features of the chapter.

2.2 Notations & Conventions

Terminology

For brevity, covariates that vary within cluster are referred to as **cluster-varying covariates** while covariates that do not vary within cluster are referred to as **cluster-constant covariates**. With respect to section 2.1, $\boldsymbol{x}_{ij}$ are cluster-varying covariates while $\boldsymbol{w}_i$ are cluster-constant covariates.

Symbols

Throughout this chapter, all vectors and matrices will be denoted by bold letters or symbols. All scalar quantities are denoted by plain letters or symbols. The vector $\boldsymbol{Y}$ is used to denote an outcome or response vector. Y_{ij} is used to denote the j-th response in the i-th cluster with $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$. The size of the i-th cluster will always be assumed n_i. For example, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ is a $n_i \times 1$ outcome vector of the i-th cluster. The symbol $\boldsymbol{X}_i$ is used to denote an $n_i \times p$ matrix of cluster-varying covariates whose j-th row is $\boldsymbol{x}_{ij}^\top$. The symbol $\boldsymbol{W}_i$ is used to denote an $n_i \times q$ matrix of cluster-constant covariates whose j-th row is $\boldsymbol{w}_i^\top$.

We assume that the true model has the following mean structure

$$h(\mu_{ij}) = \lambda_i + \beta_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{w}_i^\top \boldsymbol{\gamma}$$

with link function h and where $\mu_{ij} = E(Y_{ij} | \boldsymbol{X}_i, \boldsymbol{W}_i)$. Since the emphasis is on estimation of $\boldsymbol{\beta}$,

we write it as

$$h(\mu_{ij}) = \delta_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$$

which views $\delta_i := \lambda_i + \boldsymbol{w}_i^\top \boldsymbol{\gamma}$ as a nuisance parameter. In the scheme of this paper, it is irrelevant whether information on cluster-constant covariates are available. All that matters is that information on within-cluster covariates are available.

The symbol $\odot$ will be used to denote component-wise multiplication. For example,

$$\boldsymbol{a} \odot \boldsymbol{b} = [a_1 b_1, \ \cdots, \ a_n b_n]^\top .$$

If $\boldsymbol{a}$ is a vector and s is a function defined on $\mathbb{R}$, the notation $s(\boldsymbol{a})$ denotes a vector that result from applying the function to every element of the vector $\boldsymbol{a}$:

$$s(\boldsymbol{a}) = \begin{bmatrix} s(a_1) & \cdots & s(a_n) \end{bmatrix}^\top .$$

For example, if $\boldsymbol{X}\boldsymbol{\beta}$ is a $n \times 1$ vector, then

$$\exp(\boldsymbol{X}\boldsymbol{\beta}) = \begin{bmatrix} e^{\boldsymbol{x}_1^\top \boldsymbol{\beta}} & \ldots & e^{\boldsymbol{x}_n^\top \boldsymbol{\beta}} \end{bmatrix}^\top .$$

2.3 Motivation

Model (2.1) arises in several contexts. These include

- meta-analysis
- multi-center studies
- stratified sampling designs
- repeated measures/longitudinal studies

2.3.1 Meta-Analysis

A typical meta-analysis study involves combining results from many small studies into a single study. There are various forms of meta analyses and they typically treat parameter estimates from a collection of studies as the outcome of interest. Fleiss (1993) provides an accessible

overview of the statistical basis of meta analyses. The type of meta-analysis that is relevant to this discussion involves combining the actual data from various studies. Models of the mean structure typically contain a study-specific intercept. For example, Lee et al. (2009) discussed combining 2×2 tables from various studies and fitting a logistic regression model with random intercepts.

2.3.2 Multi-center Studies

Multi-center studies offer a convenient study design for examining the association between treatment, exposure, and disease outcomes. Analyses based on pooling of data assume that patients from all centers come from the same distribution. In practice, this may not be true. To account for heterogeneity due to centers, a center-specific intercepts are added to the mean structure. Localio et al. (2001) provided a broad overview of multi-center studies.

2.3.3 Stratified Sampling Designs

In stratified designs, a population is partitioned into a fixed number of disjoint strata based on a single variable or cross-classification of multiple variables that are thought to be related to the outcome of interest. Sampling is performed independently within each stratum. For example, matched-pair studies can be viewed as stratified data (Jewell, 2004) where strata are defined by the variable(s) used in the matching. Stratum-specific intercepts become part of the model for the mean structure and are viewed as fixed effects .

2.3.4 Repeated Measures/Longitudinal Studies (RML)

RML studies are typically used to study changes in the outcome of interest over the measurement occasions as a function of a set of covariates. Because of biological differences between individuals, we do not expect individuals with the same covariates to have the same mean structure. To account for variation between individuals, random effects are introduced into the mean structure in the form of a random intercept, random slope, or a combination thereof.

2.4 Previous Works

This section provides a survey of techniques that have been proposed for dealing with the nuisance parameter problems in general and those described in section 2.1.

2.4.1 Conditional Likelihood

For model (2.2), if we assume that δ_i is random, then maximum likelihood estimation requires that we specify a mixing distribution for δ_i. For example, we may assume that $\delta_i \sim G(\boldsymbol{\theta})$, for some G. As discussed above, if we specify an incorrect G, then estimates of $\boldsymbol{\beta}$ may be biased. To avoid specifying G, we can condition on the sufficient statistics for δ_i and obtain a conditional likelihood that is δ_i-free. Maximization of the conditional likelihood leads to consistent estimates of $\boldsymbol{\beta}$ (Neuhaus and Jewell, 1990). This method works for exponential family models with canonical links. When the link function is not canonical, then no simple sufficient statistics for δ_i exists and the method of conditional likelihood does not work.

2.4.2 Conditional GEE (CGEE)

As mentioned earlier, when δ_i is assumed random and correlated with covariates, then maximum likelihood estimates are generally biased. The between-within covariate decomposition technique was designed to avoid this problem. In doing so, this method introduces interpretation problems. Goetgeluk and Vansteelandt (2008) developed an estimating equation called conditional GEE which allows the user to ignore the mixing distribution and thus avoid the problem of dealing with random effects that are correlated with the covariates. Their method is applicable when the assumed model that generates the data is based on either the identity or log link.

Letting $\boldsymbol{V}_i = \text{var}(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{W}_i)$ and $\boldsymbol{Z}_i = \boldsymbol{V}_i^{-1}\boldsymbol{X}_i$, they constructed an unbiased estimating equation for estimating $\boldsymbol{\beta}$ under the identity link which can be expressed as follows:

$$\boldsymbol{U}_I(\boldsymbol{\beta}) = \sum_{i=1}^{K} (\boldsymbol{Z}_i - \mathbf{1}_{n_i}\bar{\boldsymbol{z}}_i^{\top})^{\top}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = \mathbf{0} \tag{2.4}$$

where

$$\bar{\boldsymbol{z}}_i = \frac{\boldsymbol{Z}_i^\top \mathbf{1}}{n_i}$$

Since $\boldsymbol{V}_i$ is unknown, they proposed estimating $\boldsymbol{V}_i$ by first fitting the random effects model (2.2) and then setting $\widehat{\boldsymbol{V}}_i$ equal to the residual covariance matrix. This means that their estimating equation is actually

$$\tilde{\boldsymbol{U}}_I(\boldsymbol{\beta}) = \sum_{i=1}^{K} (\tilde{\boldsymbol{Z}}_i - \mathbf{1}_{n_i} \bar{\tilde{\boldsymbol{z}}}_i^\top)^\top (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) = \mathbf{0} \tag{2.5}$$

where

$$\bar{\tilde{\boldsymbol{z}}}_i = \frac{\tilde{\boldsymbol{Z}}_i^\top \mathbf{1}}{n_i} \quad \text{and} \quad \tilde{\boldsymbol{Z}}_i = \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \ .$$

This raises the following questions:

- Is (2.5) still unbiased when $\boldsymbol{V}_i$ is replaced by $\widehat{\boldsymbol{V}}_i$? If not, what is the extent of the bias? This issue is not addressed.
- Are efficiency arguments given for (2.4) still valid in (2.5) given that $\widehat{\boldsymbol{V}}_i$ was obtained from a mixed model whose random effects are correlated with covariates? The results of Neuhaus et al. (2006) suggest that $\widehat{\boldsymbol{V}}_i$ is biased in this setting. Simulation studies provided by Goetgeluk and Vansteelandt (2008) do not address this situation.

Under the log link, they proposed the estimating equation

$$\boldsymbol{U}_L(\boldsymbol{\beta}) = \sum_{i=1}^{K} (\boldsymbol{Z}_i - \mathbf{1} \bar{\boldsymbol{z}}_i^\top)^\top \boldsymbol{Y}_{i,\beta} = \mathbf{0} \tag{2.6}$$

where $\boldsymbol{Y}_{i,\beta} = \boldsymbol{Y}_i \odot \boldsymbol{e}^{-\boldsymbol{X}_i \boldsymbol{\beta}}$, $\boldsymbol{Z}_i = \bar{Y}_{i,\beta} \boldsymbol{V}_i^{-1} \boldsymbol{X}_i$, $\boldsymbol{V}_i = \mathrm{var}(\boldsymbol{Y}_{i,\beta} | \boldsymbol{X}_i, \boldsymbol{W}_i)$, and

$$\bar{Y}_{i,\beta} = \frac{\boldsymbol{Y}_{i,\beta}^\top \mathbf{1}}{n_i} \ , \quad \bar{\boldsymbol{z}}_i = \frac{\boldsymbol{Z}_i^\top \mathbf{1}}{n_i} = \bar{Y}_{i,\beta} \frac{\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \mathbf{1}}{n_i} \ .$$

As with the identity link, the residual covariance $\boldsymbol{V}_i$ is assumed to be constant across clusters. Since $\boldsymbol{V}_i$ is unknown, they proposed solving the following estimating equation for $\boldsymbol{\beta}$ instead of (2.6):

$$\tilde{\boldsymbol{U}}_L(\boldsymbol{\beta}) = \sum_{i=1}^{K} (\tilde{\boldsymbol{Z}}_i - \mathbf{1} \bar{\tilde{\boldsymbol{z}}}_i^\top)^\top \boldsymbol{Y}_{i,\beta} = \mathbf{0} \tag{2.7}$$

where

$$\tilde{\boldsymbol{Z}}_i = \bar{Y}_{i,\beta}\widehat{\boldsymbol{V}}_i^{-1}\boldsymbol{X}_i \quad \text{and} \quad \bar{\tilde{\boldsymbol{z}}}_i = \frac{\tilde{\boldsymbol{Z}}_i^\top \mathbf{1}}{n_i} \ .$$

They proposed estimating $\boldsymbol{V}_i$ in the following way:

1. Obtain a preliminary estimate of $\boldsymbol{\beta}$ by solving

$$U_L(\boldsymbol{\beta}) = \sum_{i=1}^{K} (\boldsymbol{X}_i - \mathbf{1}\bar{\boldsymbol{x}}_i^\top)^\top \boldsymbol{Y}_{i,\beta} = \mathbf{0}$$

   where $\bar{\boldsymbol{x}}_i$ is the simple average of $\boldsymbol{X}_i$. Call this estimate $\widehat{\boldsymbol{\beta}}_0$

2. Compute the i-th transformed response vector by

$$\boldsymbol{Y}_{i,\beta_0} = \boldsymbol{Y}_i \odot e^{\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_0}$$

3. Fit a random effects model on the transformed response. It isn't clear what this "random effects model" is but we interpret it as the mixed model

$$E[\boldsymbol{Y}_{i,\beta_0}] = b_i + \alpha$$

   where $b_i \sim N(0, \sigma_b^2)$.

4. Estimate $\boldsymbol{V}_i$ by

$$\widehat{\boldsymbol{V}}_i = \frac{\sum_{i=1}^{K} (\boldsymbol{Y}_{i,\beta_0} - e^{\mathbf{1}\hat{\alpha}})(\boldsymbol{Y}_{i,\beta_0} - e^{\mathbf{1}\hat{\alpha}})^\top}{K}$$

5. Solve (2.7) for $\widehat{\boldsymbol{\beta}}$ using $\widehat{\boldsymbol{V}}_i$ obtained in step 4.

As with the identity link, this procedure raises the following questions:

- Is (2.7) still unbiased since $\boldsymbol{V}_i$ is replaced by $\widehat{\boldsymbol{V}}_i$?
- Are efficiency arguments given for (2.6) still valid in (2.7)?
- Is $\widehat{\boldsymbol{V}}_i$ consistent when the random effects are correlated with the covariates?

2.4.3 Optimally-Weighted Estimating Functions

Liang and Zeger (1995) discussed how nuisance parameters can break down the machinery of estimating functions. They proposed the following technique for constructing an estimating function in the presence of nuisance parameters. Define $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)^\top$. Assume that it is possible to obtain a set of estimating functions, $\{\boldsymbol{U}_i(\boldsymbol{\beta}) : i = 1, \ldots, K\}$ such that $E[\boldsymbol{U}_i(\boldsymbol{\beta})|\boldsymbol{X}_i; \boldsymbol{\beta}, \boldsymbol{\delta}] = \mathbf{0}$. Assume also that $\text{cov}(\boldsymbol{U}_i, \boldsymbol{U}_{i'}) = \mathbf{0}$. Construct an optimally-weighted estimating function $\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta})$ as follows:

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{K} E\left[\frac{\partial \boldsymbol{U}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|\boldsymbol{X}_i\right]^\top \text{var}[\boldsymbol{U}_i(\boldsymbol{\beta})|\boldsymbol{X}_i]^{-1}\boldsymbol{U}_i(\boldsymbol{\beta}) \ . \tag{2.8}$$

The nuisance parameters are embedded in the weights

$$E\left[\frac{\partial \boldsymbol{U}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|\boldsymbol{X}_i\right]^\top \text{var}[\boldsymbol{U}_i(\boldsymbol{\beta})|\boldsymbol{X}_i]^{-1} \ .$$

By constructing the estimating function this way, the role of the nuisance parameters is reduced to that of weights. Liang and Zeger argued that while the weighted estimating function depends on the nuisance parameters, its usage is recommended for the following reasons:

- $E[\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}^*); \boldsymbol{\beta}, \boldsymbol{\delta}] = \mathbf{0}$ for all $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and $\boldsymbol{\delta}^*$. This property states that unbiasedness holds even when $\boldsymbol{\delta}$ is fixed at $\boldsymbol{\delta}^*$ – an incorrect value of $\boldsymbol{\delta}$.
- $E[\boldsymbol{U}(\boldsymbol{\beta}, \tilde{\boldsymbol{\delta}}); \boldsymbol{\beta}, \boldsymbol{\delta}] = \mathbf{0}$ for all $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, and $\tilde{\boldsymbol{\delta}}$ where $\tilde{\boldsymbol{\delta}}$ is any $\sqrt{K}$-consistent estimator of $\boldsymbol{\delta}$.
- $\text{cov}[\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}), \partial \log f(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{\delta})/\partial \boldsymbol{\delta}] = \mathbf{0}$ for all $\boldsymbol{\beta}$, $\boldsymbol{\delta}$. This property implies that if it is possible to obtain a maximum likelihood estimate of $\boldsymbol{\delta}$ for fixed $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\delta}}(\boldsymbol{\beta})$, then $E[\boldsymbol{U}(\boldsymbol{\beta}, \widehat{\boldsymbol{\delta}}); \boldsymbol{\beta}, \boldsymbol{\delta}] = \mathbf{0}$ for all $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$.

These three properties are known as orthogonality properties and are enjoyed by Lindsay's (Lindsay, 1982) conditional score function. An implication of orthogonality is that the impact of the nuisance parameters is small *provided that estimates of the nuisance parameters are consistent*. This was done in GEE (Liang and Zeger, 1986) where the regression parameters were the parameter of interest while the correlation parameters were the nuisance parameters.

In the context of (2.2), this recommendation may not be useful for the following reasons:

- it may be difficult to obtain $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_K$ that are nuisance-free
- it is difficult (if not impossible) to obtain consistent estimates of δ_i, $i = 1, \ldots, K$
- efficiency statements cannot be made about $\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta})$ when $\widehat{\boldsymbol{\delta}}$ is substituted for $\boldsymbol{\delta}$.

2.4.4 Projected Estimating Function Method

Rathouz and Liang (1999) were interested in fitting the model $h(\mu_{ij}) = \delta_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$ for a matched pair design. This is exactly the same model as (2.2) except that, here, they assumed that observations within each pair are conditionally independent. They treated each pair as a stratum and viewed each δ_i as a fixed parameter that contains information on the i-th pair. To make inferences on $\boldsymbol{\beta}$ in the presence of $\boldsymbol{\delta}$, they extended the projected score method of Waterman and Lindsay (1996), developed under fully parametric settings, to the quasi-score setting. The idea is to obtain locally ancillary estimating functions when it is not possible to obtain globally ancillary estimating functions. In particular, they provided an algorithm for obtaining a second-order locally ancillary estimating function for estimating $\boldsymbol{\beta}$ and argued that second-order ancillarity is enough for practical purposes.

In their terminology, an estimating function $\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta})$ for estimating $\boldsymbol{\beta}$ is globally ancillary if

$$E[\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}^*); \boldsymbol{\beta}, \boldsymbol{\delta}] = \mathbf{0} \quad \forall\, \boldsymbol{\beta}, \boldsymbol{\delta}^*, \boldsymbol{\delta}\ .$$

Note that this is one of the orthogonality properties of Liang and Zeger (1995) described in the previous section. An alternative to global ancillarity is local ancillarity which is defined by Small and McLeish (1994) as follows. An estimating function $\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta})$ is r-th order locally ancillary if

$$b_k[\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta})] := \frac{\partial^k}{\partial \boldsymbol{\delta}^k} E\Big[\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}^*); \boldsymbol{\beta}, \boldsymbol{\delta}\Big]_{\boldsymbol{\delta}=\boldsymbol{\delta}^*} = \mathbf{0} \quad k = 0, \ldots, r$$

Under certain conditions, r-th order locally ancillarity is equivalent to

$$E[\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}^*); \boldsymbol{\delta}] = o\big\{\|\boldsymbol{\delta}^* - \boldsymbol{\delta}\|^r\big\}$$

Written this way, we may view globally ancillarity as a limiting form of locally ancillarity (Rathouz and Liang, 1999).

Consider a simple parametric case where θ is the parameter of interest and ϕ is the nuisance parameter. Waterman and Lindsay (1996) approximated the space that contains information about ϕ using the span of a Bhattacharyya basis $\boldsymbol{V}_r$. Here, $\boldsymbol{V}_r := (V_1, \ldots, V_r)^\top$ where

$$V_k = \frac{\partial^k f(y;\theta,\phi)/\partial\phi^k}{f(y;\theta,\phi)} \ , \quad k = 1, \ldots, r \ .$$

If $u(\theta,\phi)$ is the score function for estimating θ, they obtained an r-th order locally ancillary estimating function $u_a(\theta,\phi)$ as follows:

$$u_a(\theta,\phi) = u(\theta,\phi) - \Pi_{\boldsymbol{V}_r} u(\theta,\phi)$$

where $\Pi_{\boldsymbol{V}_r} u(\theta,\phi)$ is the projection of $u(\theta,\phi)$ onto the span of $\boldsymbol{V}_r$. u_a is said to be maximally correlated with u and uncorrelated with the ϕ scores - the vectors that make up $\boldsymbol{V}_r$. Intuitively, this is equivalent to saying that u_a is sensitive to θ but insensitive to ϕ.

In our setup, Rathouz and Liang extended this idea to the quasi-score setting by replacing the Bhattacharyya basis with derivatives of nuisance estimating functions rather than derivatives of nuisance score functions. They argued that for a second-order locally ancillary estimating function $\boldsymbol{U}_a(\boldsymbol{\beta},\boldsymbol{\delta})$ for estimating $\boldsymbol{\beta}$,

$$E[\boldsymbol{U}_a(\boldsymbol{\beta},\widehat{\boldsymbol{\delta}})] \approx \mathbf{0}$$

as $K \to \infty$. By virtue of insensitivity, this approach does not require $\widehat{\boldsymbol{\delta}}$ to be consistent.

A difficulty with this approach is that the estimating function can be very complicated. Furthermore, even though $\widehat{\boldsymbol{\delta}}$ does not have to be consistent, it still has to be estimated. In addition, the estimated asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ depends on $\widehat{\boldsymbol{\delta}}$. Since $\widehat{\boldsymbol{\delta}}$ can be biased, it is not clear what impact this has on the estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$.

2.4.5 Cox-Reid Adjusted Profile Estimating Function Method

When the dimension of the nuisance parameter is fixed, Wang and Hanfelt (2003) developed a theory for adjusting an estimating function so that it achieves approximate unbiasedness. Their theory is a direct extension of the adjusted profile likelihood developed by Cox and Reid (1987). Using the notation of Wang and Hanfelt, let θ denote a scalar parameter of interest and $\boldsymbol{\lambda}$ a nuisance parameter vector of fixed dimension. Then, the Cox-Reid adjusted profile likelihood is

$$\ell_{adj} = \ell(\theta, \boldsymbol{\lambda}) - \frac{1}{2} \log |-\ell_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\theta, \boldsymbol{\lambda})|$$

where $\ell_{\boldsymbol{\lambda}\boldsymbol{\lambda}}$ is the second derivative of the log likelihood $\ell(\theta, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$. This implies

$$u_{adj}(\theta, \boldsymbol{\lambda}) := u(\theta, \boldsymbol{\lambda}) - \frac{1}{2} \operatorname{trace} \left[\{ \boldsymbol{v}_{\boldsymbol{\lambda}}(\theta, \boldsymbol{\lambda}) \}^{-1} u_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\theta, \boldsymbol{\lambda}) \right] \tag{2.9}$$

where $\boldsymbol{v}(\theta, \boldsymbol{\lambda})$ is the $\boldsymbol{\lambda}$-score, $u(\theta, \boldsymbol{\lambda})$ is the θ-score, $\boldsymbol{v}_{\boldsymbol{\lambda}}(\theta, \boldsymbol{\lambda})$ is the derivative of $\boldsymbol{v}$ with respect to $\boldsymbol{\lambda}$, and $u_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\theta, \boldsymbol{\lambda})$ is the second derivative of $u(\theta, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$. Cox and Reid showed that $u_{adj}(\theta, \hat{\boldsymbol{\lambda}})$ achieves a first order plug-in bias of $O(n^{-1})$. Note that n is the sample size under the setup of independent data.

Now, let $\boldsymbol{U}(\theta, \boldsymbol{\lambda})$ be a θ-estimating function and $\boldsymbol{S}(\theta, \boldsymbol{\lambda})$ be a $\boldsymbol{\lambda}$-estimating function. Wang and Hanfelt (2003, Theorem 1) extended (2.9) to the quasi-score setting and showed that

$$\boldsymbol{U}_{adj}(\theta, \hat{\boldsymbol{\lambda}}_\theta) := \boldsymbol{U}(\theta, \hat{\boldsymbol{\lambda}}_\theta) - \frac{1}{2} \operatorname{trace} \Big\{ [\boldsymbol{S}_{\boldsymbol{\lambda}}(\theta, \boldsymbol{\lambda})]^{-1} \boldsymbol{U}_{\boldsymbol{\lambda},\boldsymbol{\lambda}}(\theta, \boldsymbol{\lambda}) \Big\}_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}_\theta}$$

achieves first-order bias of $O(K^{-1})$ provided these conditions are met:

- Orthogonality: $E[\boldsymbol{U}_{\boldsymbol{\lambda}}(\theta, \boldsymbol{\lambda})] = \boldsymbol{0}$.
- Information-unbiasedness of $\boldsymbol{S}$: $E[\boldsymbol{S}^\top \boldsymbol{S} + \boldsymbol{S}_{\boldsymbol{\lambda}}] = \boldsymbol{0}$
- 3rd Bartlett identity: $E[\boldsymbol{S}^\top \boldsymbol{U}_{\boldsymbol{\lambda}} + \boldsymbol{U}_{\boldsymbol{\lambda},\boldsymbol{\lambda}}] = \boldsymbol{0}$

They show through a series of propositions how to construct $\boldsymbol{U}(\theta, \boldsymbol{\lambda})$ and $\boldsymbol{S}(\theta, \boldsymbol{\lambda})$ that satisfies these three conditions.

There are several shortcomings of this approach.

- First, as we pointed out earlier, this method requires that the dimension of $\boldsymbol{\lambda}$ is fixed. Hence, their theory does not support the setup of model (2.2) where the number of nuisance parameters grows with the sample size.

- Second, the theory was developed under the assumption that θ is a scalar quantity. The authors commented that when $\boldsymbol{\theta}$ is a vector quantity, the adjustments can be performed separately for every element of $\boldsymbol{U}(\boldsymbol{\theta}, \boldsymbol{\lambda})$. As shown in the appendix, this task is not as easy as it sounds.

- As with previously discussed methods, the adjustment terms require taking derivatives of vector valued functions. This can lead to very complicated adjusted profile estimating functions which can make the root-finding task unfeasible.

- Estimation requires alternating between estimating the nuisance parameters and the parameter of interest.

2.4.6 Barndorf-Nielsen Profile Estimating Function Method

Severini (2002) considered the nuisance parameter problem in the setup of a scalar parameter of interest, denoted by θ, and a vector nuisance parameter, denoted by $\boldsymbol{\lambda}$. The dimension of $\boldsymbol{\lambda}$ is assumed fixed. Severini's idea was to extend Barndorff-Nielsen's (Barndorff-Nielsen, 1983) adjusted profile score to estimating functions. The idea can be described as follows. First, start with Barndorff-Nielsen's profile log-likelihood:

$$\ell_{BN}(\theta) := \ell_p(\theta) + \log \frac{|-\ell_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\theta, \hat{\boldsymbol{\lambda}}_\theta)|}{|\ell_{\boldsymbol{\lambda};\hat{\boldsymbol{\lambda}}_\theta}(\theta, \hat{\boldsymbol{\lambda}}_\theta)|} \tag{2.10}$$

where $\ell_p(\theta) = \ell(\theta, \hat{\boldsymbol{\lambda}}_\theta)$, and subscripts with respect to parameters denote derivatives with respect to those parameters. Except in a few cases, the quantity $\ell_{\boldsymbol{\lambda};\hat{\boldsymbol{\lambda}}_\theta}(\theta, \hat{\boldsymbol{\lambda}}_\theta)$ is difficult to compute. An approximation to this quantity is available (Severini, 1998):

$$I(\theta, \boldsymbol{\lambda}; \theta_0, \boldsymbol{\lambda}_0) := E_0[\boldsymbol{s}(\theta, \boldsymbol{\lambda})\boldsymbol{s}^\top(\theta, \boldsymbol{\lambda})] \tag{2.11}$$

where $\boldsymbol{s}$ is the $\boldsymbol{\lambda}$-score and E_0 denotes expectation with respect to a distribution function having parameters θ_0 and $\boldsymbol{\lambda}_0$. Expression (2.11) can be approximated by $I(\theta, \hat{\boldsymbol{\lambda}}_\theta; \hat{\theta}, \hat{\boldsymbol{\lambda}})$ where $\hat{\theta}$ and $\hat{\boldsymbol{\lambda}}$ maximizes the log-likelihood $\ell(\theta, \boldsymbol{\lambda})$. Substituting this into (2.10) gives Severini's modified profile likelihood:

$$u_M(\theta) := u(\theta, \hat{\boldsymbol{\lambda}}_\theta) + \frac{\partial}{\partial \theta} \log \frac{|-\boldsymbol{s}_{\boldsymbol{\lambda}}(\theta, \hat{\boldsymbol{\lambda}}_\theta)|}{|I(\theta, \hat{\boldsymbol{\lambda}}_\theta; \hat{\theta}, \hat{\boldsymbol{\lambda}})|} \tag{2.12}$$

where $\boldsymbol{s}_{\boldsymbol{\lambda}}$ is the derivative of the $\boldsymbol{\lambda}$-score with respect to $\boldsymbol{\lambda}$. He showed that $E[u_M(\theta)] = O(n^{-1})$ provided that the second Bartlett identity holds:

$$E[\boldsymbol{s}_{\boldsymbol{\lambda}}(\theta, \boldsymbol{\lambda}) + \boldsymbol{s}(\theta, \boldsymbol{\lambda}) \boldsymbol{s}^\top(\theta, \boldsymbol{\lambda})] = \mathbf{0}. \tag{2.13}$$

When this condition is not met, $E[u_M(\theta)] = O(1)$. To make $E[u_M(\theta)] = O(n^{-1})$ when (2.13) is not satisfied, Severini added an extra adjustment term to (2.12):

$$\begin{aligned} u_{M,adj}(\theta) = u(\theta, \hat{\boldsymbol{\lambda}}_\theta) &+ \frac{1}{2} \operatorname{trace}\Big\{ \boldsymbol{D} \boldsymbol{s}_{\boldsymbol{\lambda}}(\theta, \hat{\boldsymbol{\lambda}}_\theta)^{-1} \boldsymbol{s}_{\boldsymbol{\lambda},\theta}(\theta, \hat{\boldsymbol{\lambda}}_\theta) \Big\} \\ &- \operatorname{trace}\left\{ \boldsymbol{D} I(\theta, \hat{\boldsymbol{\lambda}}_\theta; \hat{\theta}, \hat{\boldsymbol{\lambda}})^{-1} \frac{\partial}{\partial \theta} I(\theta, \hat{\boldsymbol{\lambda}}_\theta; \hat{\theta}, \hat{\boldsymbol{\lambda}}) \right\} \end{aligned} \tag{2.14}$$

where

$$\boldsymbol{D} = -\boldsymbol{s}_{\boldsymbol{\lambda}}(\hat{\theta}, \hat{\boldsymbol{\lambda}})^{-1} I(\hat{\theta}, \hat{\boldsymbol{\lambda}}; \hat{\theta}, \hat{\boldsymbol{\lambda}})$$

For the case when no parametric model is specified, let $U(\theta, \boldsymbol{\lambda})$ denote the θ-estimating function and $\boldsymbol{S}(\theta, \boldsymbol{\lambda})$ denote the $\boldsymbol{\lambda}$-estimating function. Severini showed that (2.14) can be extended to the estimating function setting by substituting U and $\boldsymbol{S}$ for u and $\boldsymbol{s}$. The resulting adjusted estimating function satisfies $E[U_{M,adj}(\theta)] = O(K^{-1})$ provided that $U_{\boldsymbol{\lambda}} = \boldsymbol{S}_\theta$ or, at least, the equality holds approximately.

This method of adjustment is similar to those provided by Wang and Hanfelt (2003). In the semi-parametric setting, derivatives of estimating functions can lead to complicated expressions that are difficult to solve. In general, this method suffers from the same difficulties as the method of Wang and Hanfelt (2003). Furthermore, Severini provided no guidance on how to pick U and $\boldsymbol{S}$ that satisfy $U_{\boldsymbol{\lambda}} = \boldsymbol{S}_\theta$.

2.4.7 Summary

Our motivation for this chapter is how to estimate $\boldsymbol{\beta}$ in the context of model (2.2). Although many contributions have been made with respect to this model, few have satisfactorily addressed the following complications:

- lack of consistency of $\widehat{\boldsymbol{\beta}}$ caused by misspecification of the mixing distribution when δ_i is viewed as a random effect (c.f. Heagerty and Kurland, 2001)
- lack of consistency of $\widehat{\boldsymbol{\beta}}$ when δ_i is viewed as a random effect that is correlated with the covariates in the regression model (c.f. Neuhaus and McCulloch, 2006)
- when δ_i is viewed as fixed, we have a generalization of the Neyman-Scott problem. It is not clear that GEE can provide consistent estimates of $\boldsymbol{\beta}$ and δ_i, $i = 1, \ldots, K$

Under the random effects setup, conditional likelihood and CGEE can be used to circumvent the first two complications. However, these two methods are unsatisfactory in the following way. Conditional likelihood works well for exponential family models whose link is canonical to the variance function but does not work otherwise. Unbiasedness and optimality of CGEE is in question when $\boldsymbol{V}$ is replaced by $\widehat{\boldsymbol{V}}$ which is estimated from a mixed model. It is not clear what this mixed model is. Of course if assumptions of the mixed models are violated, $\widehat{\boldsymbol{V}}$ is expected to be bias.

Several estimating function methods have been proposed for addressing the third complication. These include optimally weighted estimating functions (Liang and Zeger, 1995), projected estimating functions (Rathouz and Liang, 1999), Cox-Reid type adjusted profile estimating functions (Wang and Hanfelt, 2003), and Barndorff-Nielsen type adjusted profile estimating functions (Severini, 2002). These methods share several drawbacks. First, all require estimation of the nuisance parameters. This implies that estimation procedures must alternate between estimating the nuisance parameters and the parameter of interest. In the setup of (2.2), estimates of the nuisance parameters may not be consistent since the number of nuisance parameters is growing with the sample size. Second, the efficiency matrix depends on $\boldsymbol{\delta}$. A reasonable question is whether the efficiency arguments made in favor of using these estimating

functions depend on the choice of $\widehat{\boldsymbol{\delta}}$ that is substituted for $\boldsymbol{\delta}$. Third, these estimating functions can be complicated, thus rendering the task of root-finding difficult. For example, depending on the complexity of $\boldsymbol{U}_i(\boldsymbol{\beta})$, the expression for the weights in Liang and Zeger's (1995) proposal, can lead to a very complicated estimating function whose root is difficult to find. Both the Cox-Reid estimating functions and Barndorff-Nielsen estimating functions also contain complicated derivatives.

Reiterating what we enumerated at the end of section 2.1, the goal of this chapter is to develop estimating functions for estimating $\boldsymbol{\beta}$ in the context of the cluster-specific model given in (2.2) where $\boldsymbol{\beta}$ is the parameter of interest and $\boldsymbol{\delta}$ is the nuisance parameter. We want our estimating functions to enjoy the following properties. First, we want our estimating function to be simple to implement. Second, since we cannot hope to obtain consistent estimates of the nuisance parameters, we want our estimating function to be nuisance-free. This obviates us from having to estimate $\boldsymbol{\delta}$ and substituting the inconsistent estimates into the $\boldsymbol{\beta}$-estimating function. In the context of GLMMs, this implies that we do not have to specify a mixing distribution. This is beneficial in the following ways: (1) we do not have to deal with the situation where the random effects are correlated with the covariates, and (2) we do not have to address the situation where we specify an incorrect mixing distribution that is potentially bias-inducing. We will show that this is possible when h is either the log or identity link regardless of whether we treat $\boldsymbol{\delta}$ as random or fixed.

2.5 Estimating Functions in the Presence of Cluster-Specific Nuisance Parameters: A Motivation

For convenience, we reproduce model (2.2):

$$h(\mu_{ij}) = \delta_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} \ , \quad i = 1, \ldots, K; \quad j = 1, \ldots, n_i \ . \tag{2.15}$$

where i indexes the clusters and j indexes observations within clusters. Suppose that δ_i is viewed as fixed and assume that $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$, $i = 1, \ldots, K$ is a random sample from some unknown distribution with mean function $\mu_{ij} := E(Y_{ij}|\boldsymbol{X}_i)$ specified by (2.15) and variance-covariance specified by $\boldsymbol{V}_i := \text{var}(\boldsymbol{Y}_i|\boldsymbol{X}_i)$. For the moment, assume that δ_i is known. Then, in the GEE setup, the optimal $\boldsymbol{\beta}$-estimating function is

$$\boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\delta}) = \sum_{i=1}^{K} \left[\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right]^\top \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \tag{2.16}$$

where

$$\boldsymbol{\mu}_i = [\mu_{i1} \quad \cdots \quad \mu_{in_i}]^\top \quad \text{and} \quad \boldsymbol{Y}_i = [Y_{i1} \quad \cdots \quad Y_{in_i}]^\top \ .$$

The notation $\boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\delta})$ indicates that $\boldsymbol{\delta}$ is known. If we assume an independence working correlation matrix, then (2.16) reduces to

$$\boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\delta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left[\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}} \right]^\top \left(\frac{y_{ij} - \mu_{ij}}{v_{ij}} \right) \tag{2.17}$$

where v_{ij} is the variance function. For example, we can specify $v_{ij} = \mu_{ij}(1 - \mu_{ij})$ when Y_{ij} is binary. When h is canonical to the variance function, expression (2.17) simplifies even further:

$$\boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\delta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} (y_{ij} - \mu_{ij}) \tag{2.18}$$

In practice, $\boldsymbol{\delta}$ is unknown. Rewrite (2.18) as follows to account for the fact that $\boldsymbol{\delta}$ is unknown:

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \Big(y_{ij} - \mu_{ij}(\boldsymbol{\beta}, \delta_i) \Big) \tag{2.19}$$

The notation $\mu_{ij}(\boldsymbol{\beta}, \delta_i)$ emphasizes the fact that the mean is a function of both the parameter of interest and the nuisance parameters. Note the difference between (2.18) and (2.19) is that one uses a comma and the other uses a semi-colon. The usual procedure in this situation is to estimate δ_i for fixed $\boldsymbol{\beta}$ and substitute the estimated values into (2.19) and solve

$$\boldsymbol{U}(\boldsymbol{\beta}, \widehat{\boldsymbol{\delta}}) = \mathbf{0} \ . \tag{2.20}$$

This is analogous to what is done for profile likelihoods. As with profile likelihoods, the left hand side of (2.20) suffers from plug-in bias in the sense that

$$E[\boldsymbol{U}(\boldsymbol{\beta}; \widehat{\boldsymbol{\delta}}); \boldsymbol{\beta}, \boldsymbol{\delta}] \neq \mathbf{0} \ .$$

We saw in Section 2.4 that in general the bias is $O(1)$. Note that we can rewrite (2.19) as

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i) \Big(y_{ij} - \mu_{ij}(\boldsymbol{\beta}, \delta_i) \Big) + \sum_{i=1}^{K} \bar{\boldsymbol{x}}_i \sum_{j=1}^{n_i} \Big(y_{ij} - \mu_{ij}(\boldsymbol{\beta}, \delta_i) \Big) \tag{2.21}$$

where $\bar{\boldsymbol{x}}_i$ is an average (to be defined shortly) of $\boldsymbol{X}_i$.

Suppose for the moment that we have a random sample of size n consisting of independent data (cluster size 1). From the theory of generalized linear models (McCullagh and Nelder, 1989) for independent data, we know that for canonical link models with an intercept, the following constraint holds:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\mu}_i \quad \text{or} \quad \sum_{i=1}^{n} (y_i - \hat{\mu}_i) = 0 \tag{2.22}$$

where y_i and $\hat{\mu}_i$ are scalar quantities. Motivated by this fact, we drop the second summand from (2.21) and obtain

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)\Big(y_{ij} - \mu_{ij}(\boldsymbol{\beta}, \delta_i)\Big) \ . \tag{2.23}$$

The reasoning behind this motivation is that $\mu_{ij}(\boldsymbol{\beta}, \delta_i)$ contains a cluster-specific intercept. In our setup, a constraint analogous to (2.22) should also hold:

$$\sum_{j=1}^{n_i} y_{ij} - \mu_{ij}(\widehat{\boldsymbol{\beta}}, \hat{\delta}_i) = 0 \ .$$

Both (2.19) and (2.23) contain $\boldsymbol{\delta}$ in the mean structure. In going forward, we are motivated by the following question: can we define $\bar{\boldsymbol{x}}_i$ in such a way so that (2.23) is nuisance-free? If so, are there limitations to this definition? We want our estimating function to enjoy this property so that we do not have to worry about plug-in bias. The conclusion that we came to is to define $\bar{\boldsymbol{x}}_i$ in the following way:

$$\bar{\boldsymbol{x}}_i = \frac{\boldsymbol{X}_i^\top \boldsymbol{v}_i}{\boldsymbol{1}^\top \boldsymbol{v}_i} \tag{2.24}$$

where $\boldsymbol{v}_i$ is the vector of variance function values for cluster i.

In the following sections we pursue (2.23) for the identity and log links and we assume that $\bar{\boldsymbol{x}}_i$ is defined by (2.24).

2.6 Identity Link

Consider the usual linear regression analysis setting for independent normal outcomes. Here, it is usually assumed that

$$y_i = \alpha_0 + \boldsymbol{x}_i^\top \boldsymbol{\alpha}_1 + \epsilon_i$$

where $\epsilon_i \sim N(0, \phi)$. In the language of generalized linear models, this is equivalent to saying that the variance function is 1 and the variance of Y_i has a dispersion component ϕ.

We can extend this idea to the clustered data setting by assuming that $\text{var}(\boldsymbol{Y}_i|\boldsymbol{X}_i) = \phi \boldsymbol{I}_i$, where $\boldsymbol{I}_i$ is the $n_i \times n_i$ identity matrix. This gives variance function value $v_{ij} = 1$. Under this

assumption, expression (2.23) reduces to the estimating function

$$\begin{aligned}\boldsymbol{U}(\boldsymbol{\beta}) &= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(y_{ij}-\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{K}\boldsymbol{X}_{c,i}^{\top}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}) \end{aligned} \tag{2.25}$$

where $\boldsymbol{X}_{c,i}=\boldsymbol{X}_i-\mathbf{1}\bar{\boldsymbol{x}}_i^{\top}$ is the matrix of centered covariates and $\bar{\boldsymbol{x}}_i$ is the simple average. This estimating function is not only unbiased but also nuisance-free. The unbiasedness can be seen as follows:

$$E[\boldsymbol{U}(\boldsymbol{\beta})]=\sum_{i=1}^{K}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(\delta_i+\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}-\boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta})=\sum_{i=1}^{K}\delta_i\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)=\mathbf{0}\ .$$

It is immediate from (2.25) that

$$\widehat{\boldsymbol{\beta}}=\Big(\sum_{i=1}^{K}\boldsymbol{X}_{c,i}^{\top}\boldsymbol{X}_i\Big)^{-1}\Big(\sum_{i=1}^{K}\boldsymbol{X}_{c,i}^{\top}\boldsymbol{y}_i\Big) \tag{2.26}$$

solves $\boldsymbol{U}(\widehat{\boldsymbol{\beta}})=\mathbf{0}$.

2.6.1 Asymptotic Properties of $\widehat{\boldsymbol{\beta}}$

To obtain the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$, we assume that $(\boldsymbol{Y}_i,\boldsymbol{X}_i)$, $i=1,\ldots,K$ is a random sample from some unknown distribution. This implies that

$$\boldsymbol{U}_i(\boldsymbol{\beta})=\boldsymbol{X}_{c,i}^{\top}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}),\quad i=1,\ldots,K$$

is also a random sample from some unknown distribution. Under regularity conditions and by Taylor expansion, we have

$$\begin{aligned}\mathbf{0}=\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) &\doteq \boldsymbol{U}(\boldsymbol{\beta})+\partial_{\boldsymbol{\beta}}\boldsymbol{U}(\boldsymbol{\beta}^*)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &= \sum_{i=1}^{K}\boldsymbol{U}_i(\boldsymbol{\beta})+\sum_{i=1}^{K}\partial_{\boldsymbol{\beta}}\boldsymbol{U}_i(\boldsymbol{\beta}^*)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\end{aligned}$$

Re-arranging terms gives

$$\begin{aligned}\sqrt{K}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) &= -\left\{\frac{\sum_{i=1}^{K}\partial_{\boldsymbol{\beta}}\boldsymbol{U}_i(\boldsymbol{\beta}^*)}{K}\right\}^{-1}\frac{1}{\sqrt{K}}\sum_{i=1}^{K}\boldsymbol{U}_i(\boldsymbol{\beta}) \\ &= -E\Big[\partial_{\boldsymbol{\beta}}\boldsymbol{U}_1(\boldsymbol{\beta})\Big]^{-1}\frac{1}{\sqrt{K}}\sum_{i=1}^{K}\boldsymbol{U}_i(\boldsymbol{\beta})+o_p(1)\end{aligned}$$

By the central limit theorem,

$$\frac{1}{\sqrt{K}}\sum_{i=1}^{K}\boldsymbol{U}_i(\boldsymbol{\beta}) \xrightarrow{D} N_p\Big(\mathbf{0}, E\big[\boldsymbol{U}_1(\boldsymbol{\beta})\boldsymbol{U}_1^{\top}(\boldsymbol{\beta})\big]\Big)$$

Then by Slutsky's theorem

$$\sqrt{K}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0},\boldsymbol{\Sigma})$$

where

$$\begin{aligned}\boldsymbol{\Sigma} &= E\Big[\partial_{\boldsymbol{\beta}}\boldsymbol{U}_1(\boldsymbol{\beta})\Big]^{-1}E\Big[\boldsymbol{U}_1(\boldsymbol{\beta})\boldsymbol{U}_1^{\top}(\boldsymbol{\beta})\Big]E\Big[\partial_{\boldsymbol{\beta}}\boldsymbol{U}_1(\boldsymbol{\beta})\Big]^{-1} \\ &= \lim_{K\to\infty} K\left\{\sum_{i=1}^{K}\partial_{\boldsymbol{\beta}}\boldsymbol{U}_i(\boldsymbol{\beta})\right\}^{-1}\left\{\sum_{i=1}^{K}\boldsymbol{U}_i(\boldsymbol{\beta})\boldsymbol{U}_i^{\top}(\boldsymbol{\beta})\right\}\left\{\sum_{i=1}^{K}\partial_{\boldsymbol{\beta}}\boldsymbol{U}_i(\boldsymbol{\beta})\right\}^{-1}\end{aligned}$$

Thus,

$$\operatorname{cov}(\widehat{\boldsymbol{\beta}}) \doteq \left\{\sum_{i=1}^{K}\partial_{\boldsymbol{\beta}}\boldsymbol{U}_i(\boldsymbol{\beta})\right\}^{-1}\left\{\sum_{i=1}^{K}\boldsymbol{U}_i(\boldsymbol{\beta})\boldsymbol{U}_i^{\top}(\boldsymbol{\beta})\right\}\left\{\sum_{i=1}^{K}\partial_{\boldsymbol{\beta}}\boldsymbol{U}_i(\boldsymbol{\beta})\right\}^{-1}$$

which can be consistently estimated by

$$\widehat{\operatorname{cov}}(\widehat{\boldsymbol{\beta}}) = \left\{\sum_{i=1}^{K}\partial_{\boldsymbol{\beta}}\boldsymbol{U}_i(\boldsymbol{\beta})\right\}^{-1}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\left\{\sum_{i=1}^{K}\boldsymbol{U}_i(\boldsymbol{\beta})\boldsymbol{U}_i^{\top}(\boldsymbol{\beta})\right\}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}\left\{\sum_{i=1}^{K}\partial_{\boldsymbol{\beta}}\boldsymbol{U}_i(\boldsymbol{\beta})\right\}^{-1}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}.$$

In particular,

$$\widehat{\operatorname{cov}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} \tag{2.27}$$

where $\boldsymbol{A}=\sum_{i=1}^{K}\boldsymbol{X}_{c,i}^{\top}\boldsymbol{X}_i$, and $\boldsymbol{B}=\sum_{i=1}^{K}\boldsymbol{X}_{c,i}^{\top}(\boldsymbol{y}_i-\boldsymbol{X}_i\widehat{\boldsymbol{\beta}})(\boldsymbol{y}_i-\boldsymbol{X}_i\widehat{\boldsymbol{\beta}})^{\top}\boldsymbol{X}_{c,i}$. As noted in Liang and Zeger (1986), this estimator does not require us to know the dispersion parameter even though $\operatorname{cov}(\widehat{\boldsymbol{\beta}})$ can depend on it.

2.6.2 Estimation

Since

$$\boldsymbol{X}_{c,i}^{\top}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = \boldsymbol{X}_{c,i}^{\top}(\boldsymbol{y}_i - \boldsymbol{X}_{c,i}\boldsymbol{\beta}) \ ,$$

we can think of (2.25) as the estimating function of a marginal model whose mean structure is described by a regression model without intercepts and whose covariates are the centered covariates:

$$\mu_{ij} = \boldsymbol{x}_{c,ij}^{\top}\boldsymbol{\beta}$$

This implies that existing software for GEE can be used to obtain (2.26) and (2.27). The algorithm is as follows:

1. Center all covariates that vary within cluster.
2. Specify a marginal model (without intercepts) consisting of only the centered covariates.
3. Fit the model using GEE with independence working correlation.

For example, suppose we have the following population model:

$$\mu_{ij} = \lambda_i + x_{ij}\alpha_1 + z_{ij}\alpha_2$$

where λ_i can be regarded as fixed or random. To estimate α_1 and α_2,

1. center the covariates x and z: x_c and z_c
2. pretend that $\mu_{ij} = x_{c,ij}\alpha_1 + z_{c,ij}\alpha_2$
3. estimate α_1, α_2, and their robust standard errors using GEE. In SAS (SAS Institute Inc, 2009), this can be done as follows:

```
proc genmod data=mydata;
    class id;
    model y = x_c z_c/link=identity noint;
    repeated subject=id /corr=ind; run;
```

2.6.3 Comments on Efficiency

The estimating function given in (2.25) was obtained from GEE under the assumption of independence working correlation and when the link is canonical to the variance function. As such, the estimator (2.26) is not expected to be efficient compared to some other estimator that takes the correlation into account - if such an estimator can be found. In standard GEE, we know that not much efficiency is lost if the true intra-correlation is small and you assume independence working correlation. We believe that our estimating function is well-suited for this situation and that our estimator does not give up much efficiency.

2.6.4 Situations Where This Approach is Appropriate

We believe the following situations warrant the usage of this estimating function. First, when the intra-cluster correlation is small, there is not much loss in efficiency between using an estimating function that assumes independence working correlation and an estimating function that models the intra-correlation. Second, in the GLMM setup where δ_i is assumed random and where the standard assumptions on the mixing distribution do not hold, biased estimates of the slope parameters can be avoided by making use of (2.25). Third, when δ_i is assumed fixed, GEE cannot be used to estimate $\boldsymbol{\beta}$. In this case, our estimating function is a viable alternative. Fourth, our estimating function can also be used to obtain initial estimates of slope parameters in GLMM settings where standard assumptions hold.

2.7 Log Link

In Poisson regression, we often assume that the variance function is $v_{ij} = \mu_{ij}$. In this situation, (2.23) reduces to

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)y_{ij} = \sum_{i=1}^{K}\boldsymbol{X}_{c,i}^{\top}\boldsymbol{y}_i \tag{2.28}$$

where $\boldsymbol{X}_{c.i} = \boldsymbol{X}_i - \mathbf{1}\bar{\boldsymbol{x}}_i^{\top}$,

$$\bar{\boldsymbol{x}}_i = \frac{\boldsymbol{X}_i^{\top}\boldsymbol{\zeta}_i}{\mathbf{1}^{\top}\boldsymbol{\zeta}_i} \quad \text{and} \quad \boldsymbol{\zeta}_i = \exp(\boldsymbol{X}_i\boldsymbol{\beta}) \ .$$

We can see that (2.28) is nuisance-free. Unbiasedness of (2.28) can be seen as follows:

$$E[\boldsymbol{U}(\boldsymbol{\beta})] = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)e^{\delta_i + \boldsymbol{x}_{ij}^\top\boldsymbol{\beta}} = \sum_{i=1}^{K} e^{\delta_i}\sum_{j=1}^{n_i} e^{\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i) = \mathbf{0}\ .$$

The last equality follows from the fact that for a weighted average $\bar{\boldsymbol{x}}_{w,i}$ where the weight is denoted generically as w_{ij}, the quantity $\sum_{j=1}^{n_i} w_{ij}(\boldsymbol{x}_{ij} - \bar{x}_{w,i})$ is $\mathbf{0}$. This provides a simple procedure for estimating $\boldsymbol{\beta}$. Unlike the identity link, there is no closed-form solution. As such, estimation involves iteration. A possible Newton-Rhapson approach to estimation is as follows:

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} - \left[\partial_{\boldsymbol{\beta}}\boldsymbol{U}\big(\widehat{\boldsymbol{\beta}}^{(t)}\big)\right]^{-1}\boldsymbol{U}\big(\widehat{\boldsymbol{\beta}}^{(t)}\big)$$

where

$$\partial_{\boldsymbol{\beta}}\boldsymbol{U}(\boldsymbol{\beta}) = -\sum_{i=1}^{K}\big[\partial_{\boldsymbol{\beta}}\bar{\boldsymbol{x}}_i\big]\mathbf{1}^\top\boldsymbol{y}_i \tag{2.29}$$

and

$$\partial_{\boldsymbol{\beta}}\bar{\boldsymbol{x}}_i = \frac{(\mathbf{1}^\top\boldsymbol{\zeta}_i)\boldsymbol{X}_i^\top\big(\boldsymbol{\zeta}_i \odot \boldsymbol{X}_i\big) - \boldsymbol{X}_i^\top\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top\boldsymbol{X}_i}{(\mathbf{1}^\top\boldsymbol{\zeta}_i)^2} \tag{2.30}$$

Expression (2.30) is obtained using results from Wand (2002) and Magnus and Neudecker (1999). This can be seen as follows:

$$\begin{aligned}
\partial_{\boldsymbol{\beta}}\bar{\boldsymbol{x}}_i &= \frac{\big(\mathbf{1}^\top e^{\boldsymbol{X}_i\boldsymbol{\beta}}\big)\,\partial_{\boldsymbol{\beta}}\left[\boldsymbol{X}_i^\top e^{\boldsymbol{X}_i\boldsymbol{\beta}}\right] - \boldsymbol{X}_i^\top e^{\boldsymbol{X}_i\boldsymbol{\beta}}\partial_{\boldsymbol{\beta}}\left[\mathbf{1}^\top e^{\boldsymbol{X}_i\boldsymbol{\beta}}\right]}{(\mathbf{1}^\top e^{\boldsymbol{X}_i\boldsymbol{\beta}})^2} \\
&= \frac{\big(\mathbf{1}^\top e^{\boldsymbol{X}_i\boldsymbol{\beta}}\big)\,\boldsymbol{X}_i^\top\left(\operatorname{diag} e^{\boldsymbol{X}_i\boldsymbol{\beta}}\right)\boldsymbol{X}_i - \boldsymbol{X}_i^\top e^{\boldsymbol{X}_i\boldsymbol{\beta}}\mathbf{1}^\top\left(\operatorname{diag} e^{\boldsymbol{X}_i\boldsymbol{\beta}}\right)\boldsymbol{X}_i}{(\mathbf{1}^\top e^{\boldsymbol{X}_i\boldsymbol{\beta}})^2} \\
&= \frac{\big(\mathbf{1}^\top\boldsymbol{\zeta}_i\big)\,\boldsymbol{X}_i^\top\left(\operatorname{diag}\boldsymbol{\zeta}_i\right)\boldsymbol{X}_i - \boldsymbol{X}_i^\top\boldsymbol{\zeta}_i\mathbf{1}^\top\left(\operatorname{diag}\boldsymbol{\zeta}_i\right)\boldsymbol{X}_i}{(\mathbf{1}^\top\boldsymbol{\zeta}_i)^2} \\
&= \frac{\big(\mathbf{1}^\top\boldsymbol{\zeta}_i\big)\,\boldsymbol{X}_i^\top\big(\boldsymbol{\zeta}_i \odot \boldsymbol{X}_i\big) - \boldsymbol{X}_i^\top\boldsymbol{\zeta}_i\boldsymbol{\zeta}_i^\top\boldsymbol{X}_i}{(\mathbf{1}^\top\boldsymbol{\zeta}_i)^2}
\end{aligned}$$

where the last equality follows from the fact that

$$(\operatorname{diag}\boldsymbol{\zeta}_i)\boldsymbol{X}_i = \boldsymbol{\zeta}_i \odot \boldsymbol{X}_i\ .$$

2.7.1 Asymptotic Properties of $\widehat{\boldsymbol{\beta}}$

Let $\boldsymbol{U}_i(\boldsymbol{\beta}) = \boldsymbol{X}_{c,i}^\top \boldsymbol{y}_i$. As in Section 2.6.1, assume that $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$, $i = 1, \ldots, K$ is a random sample from some unknown distribution. Then $\boldsymbol{U}_i(\boldsymbol{\beta})$ is also a random sample. Using the same arguments as in Section 2.6.1, it follows that

$$\sqrt{K}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma} = E\Big[\partial_{\boldsymbol{\beta}} \boldsymbol{U}_1(\boldsymbol{\beta})\Big]^{-1} E\Big[\boldsymbol{U}_1(\boldsymbol{\beta})\boldsymbol{U}_1^\top(\boldsymbol{\beta})\Big] E\Big[\partial_{\boldsymbol{\beta}} \boldsymbol{U}_1(\boldsymbol{\beta})\Big]^{-1}$$

The asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}$ is approximately

$$\operatorname{cov}(\widehat{\boldsymbol{\beta}}) \doteq \left\{\sum_{i=1}^K \partial_{\boldsymbol{\beta}} \boldsymbol{U}_i(\boldsymbol{\beta})\right\}^{-1} \left\{\sum_{i=1}^K \boldsymbol{U}_i(\boldsymbol{\beta})\boldsymbol{U}_i^\top(\boldsymbol{\beta})\right\} \left\{\sum_{i=1}^K \partial_{\boldsymbol{\beta}} \boldsymbol{U}_i(\boldsymbol{\beta})\right\}^{-1},$$

which can be consistently estimated by

$$\widehat{\operatorname{cov}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1} \tag{2.31}$$

where

$$\begin{aligned}
\boldsymbol{A} &= \sum_{i=1}^K \left\{ \frac{\left(\mathbf{1}^\top \widehat{\boldsymbol{\zeta}}_i\right) \boldsymbol{X}_i^\top \left(\widehat{\boldsymbol{\zeta}}_i \odot \boldsymbol{X}_i\right) - \boldsymbol{X}_i^\top \widehat{\boldsymbol{\zeta}}_i \widehat{\boldsymbol{\zeta}}_i^\top \boldsymbol{X}_i}{(\mathbf{1}^\top \widehat{\boldsymbol{\zeta}}_i)^2} \right\} \mathbf{1}^\top \boldsymbol{y}_i \\
\boldsymbol{B} &= \sum_{i=1}^K \Big\{ \boldsymbol{X}_{c,i}^\top \boldsymbol{y}_i \boldsymbol{y}_i^\top \boldsymbol{X}_{c,i} \Big\}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} \\
\widehat{\boldsymbol{\zeta}}_i &= \exp(\boldsymbol{X}_i \widehat{\boldsymbol{\beta}})
\end{aligned}$$

2.7.2 Other Variance Functions

The estimating function (2.28) was obtained from GEE under the assumption of independence working correlation and when the link is canonical to the variance function. It turns out that regardless of the variance function, as long as the link function is the log link, (2.28) continues

to be unbiased and nuisance-free. For example, with binary data, the variance function is $v_{ij} = \mu_{ij}(1-\mu_{ij})$. If we computed $\bar{\boldsymbol{x}}_i$ based on (2.24) using this variance function and substituted it into (2.23), the resulting estimating function continues to be functionally dependent on the nuisance parameters $\boldsymbol{\delta}$. This does not fit into our goal of obtaining a nuisance-free estimating function. However, we can still use (2.28) to estimate $\boldsymbol{\beta}$. Other variance functions pertinent to positive-valued outcomes that are encountered in practice include $v_{ij} = \mu_{ij}^2$ (gamma regression) and $v_{ij} = \mu_{ij}^3$ (inverse Gaussian regression).

2.8 A Generalization

Using independence working correlation as our starting point in the GEE setup, we have, from (2.17),

$$\begin{aligned}
\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}) &= \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left[\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}} \right]^\top \left(\frac{y_{ij} - \mu_{ij}}{v_{ij}} \right) \\
&= \sum_{i=1}^{K} \sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \eta_{ij}} \cdot \frac{\partial \eta_{ij}}{\partial \boldsymbol{\beta}^\top} \left(\frac{y_{ij} - \mu_{ij}}{v_{ij}} \right) \\
&= \sum_{i=1}^{K} \sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \eta_{ij}} \cdot \frac{1}{v_{ij}} \cdot \boldsymbol{x}_{ij}(y_{ij} - \mu_{ij}) \\
&= \sum_{i=1}^{K} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} c_{ij} (y_{ij} - \mu_{ij})
\end{aligned}$$

where

$$c_{ij} = \frac{\partial \mu_{ij}}{\partial \eta_{ij}} \cdot \frac{1}{v_{ij}} \ .$$

We can obtain an expression analogous to (2.23) by writing

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i) c_{ij} (y_{ij} - \mu_{ij}) + \sum_{i=1}^{K} \bar{\boldsymbol{x}}_i \sum_{j=1}^{n_i} c_{ij} (y_{ij} - \mu_{ij}) \tag{2.32}$$

Assume that

$$\sum_{j=1}^{n_i} c_{ij} y_{ij} = \sum_{j=1}^{n_i} c_{ij} \hat{\mu}_{ij} \ .$$

We make this assumption because it holds under the setup of generalized linear models for independent data. Using this assumption, we can therefore drop the second summand from (2.32) and obtain

$$\boldsymbol{U}(\boldsymbol{\beta},\boldsymbol{\delta}) = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)c_{ij}(y_{ij} - \mu_{ij}) \tag{2.33}$$

When the link is canonical to the variance function, c_{ij} evaluates to 1 and we get (2.23) as a special case.

The question now is how to define $\bar{\boldsymbol{x}}_i$ so that (2.33) is unbiased and nuisance–free? The next two sections consider this question for the gamma variance function and inverse Gaussian variance function under the log link.

2.8.1 Log Link and Gamma Variance Function

Consider the case where $v_{ij} = \mu_{ij}^2$. In this case, $c_{ij} = 1/\mu_{ij}$. If we choose weights of $c_{ij}\mu_{ij} = 1$ to define $\bar{\boldsymbol{x}}_i$, we have

$$\bar{\boldsymbol{x}}_i = \frac{\boldsymbol{X}_i^\top \mathbf{1}}{n_i}$$

and (2.33) becomes

$$\begin{aligned}
\boldsymbol{U}(\boldsymbol{\beta},\boldsymbol{\delta}) &= \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(1/\mu_{ij})(y_{ij} - \mu_{ij}) \\
&= \sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)y_{ij}}{\mu_{ij}} \\
&= \sum_{i=1}^{K} e^{-\delta_i}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)y_{ij}e^{-\boldsymbol{x}_{ij}\boldsymbol{\beta}}
\end{aligned}$$

This estimating function still depends on the nuisance parameter. However, we can drop $e^{-\delta_i}$ and obtain

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)y_{ij}e^{-\boldsymbol{x}_{ij}\boldsymbol{\beta}} = \sum_{i=1}^{K}\boldsymbol{X}_{c,i}^\top(\boldsymbol{y}_i \odot \boldsymbol{\xi}_i) \tag{2.34}$$

where $\boldsymbol{X}_{c,i} = \boldsymbol{X}_i - \mathbf{1}\bar{\boldsymbol{x}}_i^\top$ and $\boldsymbol{\xi}_i = e^{-\boldsymbol{X}_i\boldsymbol{\beta}}$. Note that (2.34) is nuisance-free and unbiased:

$$E\big[\boldsymbol{U}(\boldsymbol{\beta})\big] = \sum_{i=1}^{K} e^{\delta_i} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i) = \mathbf{0}.$$

Thus, the solution to $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ is consistent with estimated asymptotic covariance given by

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}) = \left\{ \sum_{i=1}^{K} \boldsymbol{X}_i^\top (\widehat{\boldsymbol{y}}_i \odot \boldsymbol{X}_{c,i}) \right\}^{-1} \left\{ \sum_{i=1}^{K} \boldsymbol{X}_{c,i}^\top \widehat{\boldsymbol{y}}_i \widehat{\boldsymbol{y}}_i^\top \boldsymbol{X}_{c,i} \right\} \left[\left\{ \sum_{i=1}^{K} \boldsymbol{X}_i^\top (\widehat{\boldsymbol{y}}_i \odot \boldsymbol{X}_{c,i}) \right\}^{-1} \right]^\top \tag{2.35}$$

where $\widehat{\boldsymbol{y}}_i = \boldsymbol{y}_i \odot \widehat{\boldsymbol{\xi}}_i$, and $\widehat{\boldsymbol{\xi}}_i = e^{-\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}}$.

As mentioned in Section 2.7.2, the estimating function (2.28) is also unbiased for this situation. Assessing which is more asymptotically efficient requires being able to compare (2.31) with (2.35). It is not clear how to go about doing this except through simulation. On the other hand, in the finite sample setting, we can say little about efficiency except that the estimating function given in (2.34) takes into account the true variance function whereas (2.31) takes into account the true variance function only when $v_{ij} = \mu_{ij}$.

2.8.2 Log Link and Inverse Gaussian Variance Function

When $v_{ij} = \mu_{ij}^3$, we have $c_{ij} = 1/\mu_{ij}^2$. If we choose weights $c_{ij}\mu_{ij} = 1/\mu_{ij}$, then

$$\bar{\boldsymbol{x}}_i = \frac{\boldsymbol{X}_i^\top \boldsymbol{\varphi}_i}{\mathbf{1}^\top \boldsymbol{\varphi}_i}$$

where $\boldsymbol{\varphi}_i = e^{-\boldsymbol{\eta}_i}$ so that

$$\boldsymbol{U}(\boldsymbol{\beta}, \boldsymbol{\delta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(1/\mu_{ij}^2)(y_{ij} - \mu_{ij}) \ .$$

In this expression, the nuisance parameters are embedded in $\bar{\boldsymbol{x}}_i$ and μ_{ij}. If we weight the residuals by $(1/e^{-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}})$ instead of $(1/\mu_{ij}^2)$ and redefine $\bar{\boldsymbol{x}}_i$ by

$$\bar{\boldsymbol{x}}_i = \frac{\boldsymbol{X}_i^\top \boldsymbol{\xi}_i}{\mathbf{1}^\top \boldsymbol{\xi}_i}$$

where $\boldsymbol{\xi}_i = e^{-\boldsymbol{X}_i\boldsymbol{\beta}}$, then the following estimating function is nuisance-free and unbiased:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)y_{ij}e^{-\boldsymbol{x}_{ij}^\top\boldsymbol{\beta}} = \sum_{i=1}^{K}\boldsymbol{X}_{c,i}^\top\tilde{\boldsymbol{y}}_i \tag{2.36}$$

Here, $\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i \odot \boldsymbol{\xi}_i$. Under the assumption that $(\boldsymbol{Y}_i, \boldsymbol{X}_i)$, $i = 1, \ldots, K$ is a random sample, the root of (2.36) has asymptotic covariance matrix given approximately by

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}^{-1})^\top \tag{2.37}$$

where

$$\begin{aligned}
\boldsymbol{A} &= \sum_{i=1}^{K}\boldsymbol{X}_{c,i}^\top(\tilde{\boldsymbol{y}} \odot \boldsymbol{X}_i) - \sum_{i=1}^{K}(\partial_{\boldsymbol{\beta}}\bar{\boldsymbol{x}}_i)\mathbf{1}^\top\tilde{\boldsymbol{y}}_i \\
\boldsymbol{B} &= \sum_{i=1}^{K}\boldsymbol{X}_{c,i}^\top\tilde{\boldsymbol{y}}_i\tilde{\boldsymbol{y}}_i^\top\boldsymbol{X}_{c,i} \\
\partial_{\boldsymbol{\beta}}\bar{\boldsymbol{x}}_i &= -\frac{(\mathbf{1}^\top\boldsymbol{\xi}_i)\boldsymbol{X}_i^\top(\boldsymbol{\xi}_i \odot \boldsymbol{X}_i) - (\boldsymbol{X}_i^\top\boldsymbol{\xi}_i)(\boldsymbol{X}_i^\top\boldsymbol{\xi}_i)^\top}{(\mathbf{1}^\top\boldsymbol{\xi}_i)^2}
\end{aligned}$$

Expression (2.37) can be consistently estimated by replacing all instances of $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}$.

As in the previous section, the estimating function (2.28) is also unbiased for this situation. Questions of asymptotic efficiency between the root of (2.28) and the root of (2.36) can only be answered if we can compare (2.31) with (2.37). Alas, this task is not straightforward. A separate simulation study assessing the relative efficiency between the root of (2.28) and the root of (2.36) must be conducted in order to say anything meaningful. This task is left for future research and will not be addressed in this dissertation.

2.9 Connections To the Cox Model

Consider the proportional hazards model

$$\lambda_j(t) = \lambda_0(t)e^{\boldsymbol{x}_j^\top\boldsymbol{\beta}}\ , \quad j = 1, \ldots, n$$

and assume that there are r event times ordered as $t_{(1)} < \cdots < t_{(r)}$. Assuming no ties, Cox (1972) showed that the relevant likelihood for estimating $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{r} \frac{e^{\boldsymbol{x}_{(k)}^\top \boldsymbol{\beta}}}{\sum_{\ell \in \mathcal{R}(t_{(k)})} e^{\boldsymbol{x}_\ell^\top \boldsymbol{\beta}}}$$

where $t_{(k)}$ is the k-th failure time, $\boldsymbol{x}_{(k)}$ is the covariate of the subject who failed at $t_{(k)}$, and $\mathcal{R}(t_{(k)})$ is the risk set immediately prior to $t_{(k)}$. This can be rewritten as

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{n} \left\{ \frac{e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}}{\sum_{\ell \in \mathcal{R}(t_j)} e^{\boldsymbol{x}_\ell^\top \boldsymbol{\beta}}} \right\}^{\Delta_j}$$

where Δ_j is the censoring indicator for the j-th subject. If the subject is not censored then $\Delta_j = 1$, otherwise, it is 0. This implies that the score function is

$$u(\boldsymbol{\beta}) = \sum_{j=1}^{n} \left\{ \boldsymbol{x}_j - \frac{\sum_{\ell \in \mathcal{R}(t_j)} \boldsymbol{x}_\ell e^{\boldsymbol{x}_\ell^\top \boldsymbol{\beta}}}{\sum_{\ell \in \mathcal{R}(t_j)} e^{\boldsymbol{x}_\ell^\top \boldsymbol{\beta}}} \right\} \Delta_j \ . \tag{2.38}$$

Note that we can rewrite (2.38) as

$$u(\boldsymbol{\beta}) = \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_{t_j}) \Delta_j \ . \tag{2.39}$$

where

$$\bar{\boldsymbol{x}}_{t_j} = \frac{\sum_{\ell \in \mathcal{R}(t_j)} \boldsymbol{x}_\ell e^{\boldsymbol{x}_\ell^\top \boldsymbol{\beta}}}{\sum_{\ell \in \mathcal{R}(t_j)} e^{\boldsymbol{x}_\ell^\top \boldsymbol{\beta}}}$$

is the weighted average of covariates of all individuals in $\mathcal{R}(t_j)$.

The case where failures occur at the same time has connections to our estimating function given in (2.28). To see this, assume that the sample size is n_1 and that there is only one failure time, $t_1 = 1$. Since the risk set $\mathcal{R}_{t_1}$ consists of everyone in the sample, it follows that

$$\bar{\boldsymbol{x}}_1 = \frac{\sum_{j=1}^{n_1} \boldsymbol{x}_j e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}}{\sum_{j=1}^{n_1} e^{\boldsymbol{x}_j^\top \boldsymbol{\beta}}}$$

and

$$u(\boldsymbol{\beta}) = \sum_{j=1}^{n_1} (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_1)\Delta_j \tag{2.40}$$

Expression (2.40) is the score assuming that the adjustment for ties uses the Breslow (1974) method.

We can extend (2.40) to cover the situation where the proportional hazards model has a stratified component in the sense that there is a different baseline hazard component for each stratum. Assume that there are K strata and within the i-th stratum, the sample size is n_i. If for all i, there is only one failure time in the i-th stratum, denoted by $t_i = i$, then the score function is

$$u(\boldsymbol{\beta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \boldsymbol{x}_i)\Delta_{ij} \tag{2.41}$$

But this is exactly the estimating function given in (2.28) if we view the cluster index i in our setup as both a failure time and a stratum, and y_{ij} as Δ_{ij}.

This connection between the estimating function for the log link and the Cox model implies that existing software for fitting Cox models can be used to obtain the root of (2.28). In SAS (SAS Institute Inc, 2009), for example, the syntax

```
proc phreg data=mydata;
    model clusterid*y(0)= x;
    strata clusterid;
run;
```

can be used to obtain estimates of β associated with x. This can similarly be done in R (R Development Core Team, 2008) using the syntax

```
coxph( Surv(clusterid, y) ~ x + strata(clusterid), data=mydata)
```

Though these software tools can be used to obtain estimates of $\boldsymbol{\beta}$, the standard errors are incorrect since they are likelihood-based. In particular, the covariance matrix of $\widehat{\boldsymbol{\beta}}$ returned by these software is

$$\text{cov}(\widehat{\boldsymbol{\beta}}) = \left\{ \sum_{i=1}^{K} \boldsymbol{U}_i(\widehat{\boldsymbol{\beta}}) \boldsymbol{U}_i(\widehat{\boldsymbol{\beta}})^\top \right\}^{-1} .$$

2.10 Connections to the Conditional Likelihood

The covariate centering technique of removing the nuisance parameters that is used in (2.23) and (2.24) has connections to conditional likelihood for *independent* data.

2.10.1 Linear Regression

In linear regression, $Y_1, \ldots, Y_n$ are assumed independent normals with mean $\mu_i = \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}$ and variance σ^2. Recall that the sufficient statistics for the intercept is $T = \sum_{i=1}^n Y_i$. A straightforward calculation shows that conditional on $T = t$, the conditional score function is

$$\boldsymbol{u}_c(\boldsymbol{\beta}) = \sum_{i=1}^n (\boldsymbol{x}_i^\top - \bar{\boldsymbol{x}})(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) \tag{2.42}$$

where $\bar{\boldsymbol{x}}$ is the simple average. This is exactly the same as (2.25).

2.10.2 Poisson Regression

Suppose that the outcome variables $Y_1, \ldots, Y_n$ are independent Poisson counts with mean and variance $\mu_i = E[Y_i] = \text{var}(Y_i)$ where

$$\log \mu_i = \beta_0 + x_i \beta \ .$$

The sufficient statistics for the intercept is $T = \sum_i Y_i$. Given $T = t$, the conditional score function is

$$u_c(\beta) = \sum_{i=1}^n (x_i - \bar{x}) y_i \tag{2.43}$$

where $\bar{x}$ is

$$\bar{x} = \frac{\sum_{i=1}^n x_i e^{x_i \beta}}{\sum_{i=1}^n e^{x_i \beta}} \ .$$

We can see that the conditional score function is exactly the same as (2.28) if we view each independent observation as a cluster of size 1.

Derivations leading up to (2.42) and (2.43) are provided in the appendix.

2.11 Connections to De-Sensitization Methods

The projected estimating functions (Rathouz and Liang, 1999) and the adjusted profile estimating functions (Wang and Hanfelt, 2003; Severini, 2002) can be viewed as de-sensitized estimating functions. As used here, the term de-sensitized means that the estimating functions for estimating $\boldsymbol{\beta}$ are approximately unbiased even when nuisance parameters are fixed at incorrect values. The following sections connect the estimating functions developed in Sections 2.6, 2.7, and 2.8 to projected estimating functions and Cox-Reid-like estimating functions.

2.11.1 Projected Estimating Function

It turns out that for the canonical identity link and the canonical log link, the projected estimating functions developed by Rathouz and Liang are the same as the estimating functions developed in Sections 2.6 and 2.7. The following proposition formally states this assertion.

Proposition 1. Assume that $(Y_{ij}, \boldsymbol{X}_{ij})$, $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$ are independent and that $h(\mu_{ij}) = \delta_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$. Denote by $\boldsymbol{U}_0$ the optimal estimating function for estimating $\boldsymbol{\beta}$ when $\boldsymbol{\delta}$ is fixed and when the working correlation structure is assumed independent:

$$\boldsymbol{U}_0 := \boldsymbol{U}_0(\boldsymbol{\beta}; \boldsymbol{\delta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left[\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}} \right]^\top v_{ij}^{-1} (y_{ij} - \mu_{ij})$$

Let $\boldsymbol{U}_{PEF} := \boldsymbol{U}_{PEF}(\boldsymbol{\beta}; \boldsymbol{\delta})$ denote the second-order locally ancillary projected estimating function. When the link function is canonical to the variance function, then the following holds:

1. **identity link.** $\boldsymbol{U}_{PEF}$ is identical to (2.25).

2. **log link.** $\boldsymbol{U}_{PEF}$ is identical to (2.28).

The proof is provided in the appendix.

In Section 2.8.1 and 2.8.2, we considered alternatives to (2.28) when the log link is not canonical to the variance function. It turns out that the projected estimating function method produces the same estimating functions given in (2.33) if $\bar{\boldsymbol{x}}_i$ is defined by weights $c_{ij}\mu_{ij}$.

Proposition 2. Assume that $(Y_{ij}, \boldsymbol{X}_{ij})$, $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$ are independent and that $h(\mu_{ij}) = \delta_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$. Let $\boldsymbol{U}_0$ and $\boldsymbol{U}_{PEF}$ be defined as in Proposition 1. If h is the log link, then under both the gamma variance function and the inverse Gaussian variance function, $\boldsymbol{U}_{PEF}$ is identical to (2.33) provided we define $\bar{\boldsymbol{x}}_i$ with weights $c_{ij}\mu_{ij}$.

Proof. See the appendix. □

Note that $\boldsymbol{U}_{PEF}$ contains the stratum-specific nuisance parameters.

2.11.2 Cox-Reid Type Adjusted Profile Estimating Function

In the previous section, connections between our covariate-centered estimating function and the projected estimating function were established. In this section, we point out a similar connection to the adjusted profile estimating function developed by Wang and Hanfelt (2003).

Proposition 3. Assume that $(Y_{ij}, \boldsymbol{X}_{ij})$, $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$ are independent and that $h(\mu_{ij}) = \delta_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$. Denote by $\boldsymbol{U}_0$ the optimal estimating function for estimating $\boldsymbol{\beta}$ when $\boldsymbol{\delta}$ is fixed and when the working correlation is assumed independent:

$$\boldsymbol{U}_0 := \boldsymbol{U}_0(\boldsymbol{\beta}; \boldsymbol{\delta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left[\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}} \right]^\top v_{ij}^{-1} (y_{ij} - \mu_{ij})$$

Let $\boldsymbol{U}_{APrEF} := \boldsymbol{U}_{APrEF}(\boldsymbol{\beta}; \boldsymbol{\delta})$ denote the adjusted profile estimating function. When the link function is canonical to the variance function, then the following holds:

1. **identity link.** $\boldsymbol{U}_{APrEF}$ is identical to (2.25).
2. **log link.** $\boldsymbol{U}_{APrEF}$ is identical to (2.28).

The proof is provided in the appendix.

2.11.3 The Binomial Variance Function

It was commented earlier that for the log link with binomial variance function $v_{ij} = \mu_{ij}(1-\mu_{ij})$, using v_{ij} to obtain $\bar{\boldsymbol{x}}_i$ does not eliminate the nuisance parameters. However, even for this variance function, (2.28) continues to be unbiased and nuisance-free. The following proposition

shows that the projected estimating function and the adjusted profile estimating function does not eliminate the nuisance parameters. These methods lead to more complicated estimating functions which can be difficult to use.

Proposition 4. Assume that $(Y_{ij}, \boldsymbol{X}_{ij})$, $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$ are independent and that $\log(\mu_{ij}) = \delta_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$. Suppose that $Y_{ij} \sim \text{Bernoulli}(\mu_{ij})$ and that the variance function is $v_{ij} = \mu_{ij}(1 - \mu_{ij})$. Let $\boldsymbol{U}_0$ denote the optimal estimating function for estimating $\boldsymbol{\beta}$ when $\boldsymbol{\delta}$ is fixed and when the working correlation is assumed independent:

$$\boldsymbol{U}_0 := \boldsymbol{U}(\boldsymbol{\beta}; \boldsymbol{\delta}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}(1 + \psi_{ij})(y_{ij} - \mu_{ij})$$

where $\psi_{ij} = \mu_{ij}/(1-\mu_{ij})$. Similarly, for a fixed $\boldsymbol{\beta}$, the optimal estimating function for estimating δ_i is

$$h_i := h(\delta_i; \boldsymbol{\beta}) = \sum_{j=1}^{n_i} (1 + \psi_{ij})(y_{ij} - \mu_{ij}) \ .$$

Then the following holds:

1. **projected estimating function.** The projected estimating function is

$$\boldsymbol{U}_{PEF} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(1 + \psi_{ij})(y_{ij} - \mu_{ij}) \tag{2.44}$$

   where $\bar{\boldsymbol{x}}_i$ is weighted by the odds ψ_{ij}.

2. **adjusted profile estimating function.** The Cox-Reid type adjusted profile estimating function is

$$\begin{aligned} \boldsymbol{U}_{APrEF} = &\sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij}^\top - \bar{\boldsymbol{x}}_i)(1 + \psi_{ij})(y_{ij} - \mu_{ij}) \\ &+ \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \frac{(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)\psi_{ij}^2}{\mathbf{1}^\top \boldsymbol{\psi}_i} - \frac{1}{2} \sum_{i=1}^{K} \partial_{\delta_i} \bar{\boldsymbol{x}}_i \end{aligned} \tag{2.45}$$

where $\bar{\boldsymbol{x}}_i$ is weighted by the odds ψ_{ij}, and

$$\partial_{\delta_i}\bar{\boldsymbol{x}}_i = \frac{(\mathbf{1}^\top\boldsymbol{\psi}_i)\boldsymbol{X}_i^\top[\boldsymbol{\psi}_i \odot (\mathbf{1}+\boldsymbol{\psi}_i)] - \boldsymbol{X}_i^\top\boldsymbol{\psi}_i\mathbf{1}^\top[\boldsymbol{\psi}_i \odot (\mathbf{1}+\boldsymbol{\psi}_i)]}{(\mathbf{1}^\top\boldsymbol{\psi}_i)^2}$$

The proof is provided in the appendix.

Note that (2.44) and (2.45) embeds the nuisance parameters in $\bar{\boldsymbol{x}}_i$ and μ_{ij}. These functions are algebraically more complex, especially the adjusted profile estimating function. The derivatives required for Newton-Rhapson may be difficult to obtain analytically but can be remedied by numerical differentiation. However, it is not clear whether such an effort results in estimates that are more efficient than those obtained from (2.28). The reason is that the efficiency matrix depends on estimates of the nuisance parameters.

2.12 Conclusion

In this chapter of the dissertation, we consider the problem of estimating a within-cluster covariate parameter $\boldsymbol{\beta}$ in the presence of nuisance parameters in the following model:

$$h(\mu_{ij}) = \lambda_i + \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + \boldsymbol{w}_i^\top\boldsymbol{\gamma}\ , \quad i = 1,\ldots,K; \quad j = 1\ldots,n_i$$

where h is a link function and λ_i is a nuisance parameter. Here, λ_i can either be fixed or random. As K grows large, so do the number of nuisance parameters. A number of methods have been suggested in the literature for addressing this sort of problem; for example, the projected estimating function method of Rathouz and Liang (1999) and the adjusted profile estimating function method of Wang and Hanfelt (2003). These methods seek to adjust an estimating function in order to achieve unbiasedness. This centering approach does not eliminate nuisance parameters from the estimating function but, rather, make it less sensitive to fixed values of the nuisance parameters.

We do not take this approach. Through a simple procedure we obtained two unbiased estimating functions for estimating $\boldsymbol{\beta}$ that is completely free of nuisance parameters – one for the identity link and one for the log link. Solutions to our estimating function can be obtained

by using existing software for fitting linear models. For example, the connection between our estimating function with respect to the log link and the Cox model enables the use of existing software for fitting the proportional hazards model to obtain roots to our estimating function.

Our method can be thought of as a simple extension of the method of conditional likelihood for independent data. Furthermore, we have shown that under the log and identity link functions, if certain conditions are met then our estimating function equals the projected estimating function of Rathouz and Liang (1999) and the adjusted profile estimating function of Wang and Hanfelt (2003).

It is expected that our estimating function will not be optimal because it assumes working independence. At this point, it is not clear how much efficiency is lost by making this assumption. Other methods estimate the nuisance parameters and substitute these values into the efficiency matrix. Since the estimated nuisance parameters are generally inconsistent, it is not clear that these methods are more efficient than our method. This is something to be investigated in the future.

We propose the use of our estimating functions in non-standard situations. For example, if the λ_i's are considered fixed, then GEE is not applicable. However, our estimating functions provide consistent estimates of effects associated with covariates that vary within cluster provided that the link functions are either the identity or log function. Also, if λ_i is considered random but correlated with the covariates, then standard GLMMs do not apply. Our estimating functions are also applicable in this situation. In standard GLMM situations under the log or identity links, our estimating function can also be used to obtain initial estimates of the parameters associated with covariates that vary within cluster.

In the next chapter, we provide some connections between this nuisance parameter problem, outcome-dependent sampling, and our estimating functions.

Chapter 3

Biased Sampling, Clustered Binary Data, and Regression Models

3.1 Introduction

Random sampling is the preferred method of data collection for making inferences about population quantities. In practice, however, study designs can be limited by many practical constraints that render random sampling ineffective. For example, when the outcome of interest is rare, random sampling produces very few units that have the sought-after outcome. To obtain a sufficient number of these outcomes under random sampling, a very large sample size is required. This can result in costs in excess of what is budgeted for the study. Biased sampling may be needed in order to stay within budget and collect enough study participants with the outcome of interest. Also, when measuring covariates is much more expensive relative to measuring outcomes, resources are more efficiently used by disproportionately measuring the covariates of those with the desired outcomes. This induces a biased sample. While biased sampling can be operationally efficient, it introduces analytical difficulties. When these difficulties are either not addressed or addressed incorrectly, estimates based on the biased sample are biased. The source of this bias lies in the fact that, unlike random sampling, biased sampling does not produce a sample that is representative of the population.

The literature on biased sampling in connection with independent data is vast and it is not our goal to add to this. The purpose of this chapter of the dissertation is to address biased sampling in the context of correlated binary data (i.e., family data, clustered data)

which has attracted a great deal of attention recently. In particular, we want to examine the relationship between the sample model and the population model when selection of clusters is based on cluster totals (to be described in the following section). Once a relationship between the sample model and the population model is established, we inquire whether the structure of the sample model can be used advantageously to construct unbiased estimating equations for estimating certain parameters in a specified population model. Various suggestions based on the likelihood have been proposed for this design but as we will discuss below, these can be difficult to implement.

This chapter is organized as follows. Section 3.2 introduces the necessary notations. Section 3.3 provides a brief introduction to three biased sampling schemes often used to study clustered data and section 3.4 provides a summary of various methodological contributions associated with these sampling schemes. As part of this summary, we will point out some of the shortcomings of these works. This serves as the motivation for this chapter of the dissertation.

3.2 Notations & Conventions

Terminology. The definition of **biased sampling** is broad. In this dissertation, we restrict its definition to mean a sampling scheme where the intensity with which a cluster is sampled depends on the outcome vector of the cluster. Other synonyms for biased sampling include *ascertainment*, *outcome-dependent*, *response-selective*, *response-biased*, and *choice-based* sampling.

For clarity, we make the following distinction between **sample models** and **population models**: a sample model is a model that is induced by the sampling scheme whereas a population model is a model that we specify or hypothesize to describe some aspect of the distribution in the population such as the mean structure, the variance structure, or even the joint distribution of the outcomes in each cluster.

Symbols. Throughout this chapter, all vectors and matrices will be denoted by bold letters or symbols. All scalar quantities are denoted by unbolded letters or symbols. The vector $\boldsymbol{Y}$ is used to denote a response vector. Y_{ij} is used to denote the j-th response in the i-th cluster with

$i = 1, \ldots, K$ and $j = 1, \ldots, n_i$. The size of the i-th cluster is assumed to be n_i. For example, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ is a $n_i \times 1$ response vector of the i-th cluster. The total T_i in the i-th cluster is defined as $T_i = \sum_{j=1}^{n_i} Y_{ij}$.

We make the following distinctions regarding covariates. Those that vary within cluster are denoted by $\boldsymbol{x}$ and we will refer to them as **cluster-varying** covariates. Those that do not vary within cluster are denoted by $\boldsymbol{w}$ and we will refer to them as **cluster-constant** covariates. In particular, $\boldsymbol{x}_{ij}$ is the $p \times 1$ covariate vector for the j-th member in the i-th cluster while $\boldsymbol{w}_i$ is a $q \times 1$ vector of cluster-constant covariates in the i-th cluster. $\boldsymbol{X}_i$ is a $n_i \times p$ matrix whose j-th row is $\boldsymbol{x}_{ij}^\top$. Similarly, $\boldsymbol{W}_i$ is a $n_i \times q$ matrix where all of the rows are $\boldsymbol{w}_i^\top$. For convenience, we sometimes write $\boldsymbol{Z}_i = (\boldsymbol{X}_i, \boldsymbol{W}_i)$.

When discussing the proband design, a slight change of notation is necessary. The index j will start from 0 instead of 1 and the cluster size will be $n_i + 1$ instead of n_i. For example, the i-th cluster obtained from a proband design has outcome vector $\boldsymbol{Y}_i = (Y_{i0}, Y_{i1}, \ldots, Y_{in_i})^\top$ where Y_{i0} is the disease status of the proband in the i-th cluster.

Sampling Conventions. The sampling intensity function (or just sampling function) will be denoted by $\pi(\boldsymbol{Y}_i)$ if sampling depends only on $\boldsymbol{Y}_i$. It will be denoted by $\pi(\boldsymbol{Y}_i, \boldsymbol{W}_i)$ if sampling depends on both $\boldsymbol{Y}_i$ and $\boldsymbol{W}_i$.

The indicator variable S_i will be used to denote the sampling status of the i-th cluster: 1 if the i-th cluster is sampled and 0 if it is not.

Marginal Models. The following notations will be used to describe marginal models. Marginal mean structures are described by

$$h(\mu_{ij}) = \beta_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_1 + \boldsymbol{w}_i^\top \boldsymbol{\beta}_2, \quad i = 1, \ldots, K; \quad j = 1, \ldots, n_i \tag{3.1}$$

where i indexes clusters and j indexes observations within clusters and

- h is a link function
- $\boldsymbol{\beta}_1$ is a $p \times 1$ parameter vector associated with cluster-varying covariates which we some-

times refer to as the slope

- $\boldsymbol{\beta}_2$ is a $q \times 1$ parameter vector associated with cluster-constant effects.

Since our interest is in binary data, we define μ_{ij} by

$$\mu_{ij} = E[Y_{ij}|\boldsymbol{X}_i, \boldsymbol{W}_i] = \Pr(Y_{ij} = 1; \boldsymbol{x}_{ij}, \boldsymbol{w}_i)$$

and we assume that $\text{var}(\boldsymbol{Y}_i|\boldsymbol{X}_i, \boldsymbol{W}_i) = \boldsymbol{V}_i$ where $\boldsymbol{V}_i$ is a matrix of variance-covariance functions.

Generalized Linear Mixed Models (GLMMs). Generalized linear mixed models are specified as

$$h(\mu_{ij}^c) = \alpha_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_1 + \boldsymbol{w}_i^\top \boldsymbol{\alpha}_2 + \boldsymbol{u}_{ij}^\top \boldsymbol{b}_i \quad i = 1, \ldots, K;\ j = 1, \ldots, n_i \tag{3.2}$$

where i indexes clusters and j indexes observations within clusters and

- h is a link function
- α_1 is a $p \times 1$ parameter vector associated with cluster-varying covariates
- α_2 is a $q \times 1$ parameter vector associated with cluster-constant covariates
- $\boldsymbol{b}_i$ is a $r \times 1$ random effects vector associated with the i-th cluster such that $\boldsymbol{b}_i \sim G(\boldsymbol{\theta})$ for some distribution G and m-dimensional parameter $\boldsymbol{\theta}$
- $\text{cov}(\boldsymbol{b}_i, \boldsymbol{Z}_i) = \boldsymbol{0}$, where $\boldsymbol{Z}_i = [\boldsymbol{X}_i,\ \boldsymbol{W}_i]$
- $\mu_{ij}^c = E[Y_{ij}|\ \boldsymbol{b}_i, \boldsymbol{Z}_i]$ which is equal to $\Pr(Y_{ij} = 1|\boldsymbol{b}_i, \boldsymbol{z}_{ij})$ for binary data
- $\boldsymbol{u}_{ij}$ is $r \times 1$ vector of covariates

Furthermore, it is assumed that given $\boldsymbol{b}_i$, the Y_{ij}'s are independent.

3.3 Biased Sampling

Biased sampling can have many meanings. In this dissertation, biased sampling refers to the situation where the process by which a cluster ends up in the sample depends on the response vector of the cluster. In some situations, biased sampling is intentional in the sense that we actively look for clusters whose response vector has a certain pattern. In other situations, biased sampling is more subtle. In biomedical studies, for example, when patients are selected for study, patients with long survival times are over-represented in the sample. Analyses based on the sample generally overestimates survival times. This is known as length-biased sampling; $\pi(y_i) \propto y_i$.

Davidov and Zelen (2001) noted that large families have a greater probability than smaller families of having at least one diseased member. This implies that the distribution of family size in referent registries emphasizes larger families than in the population. They showed that analyses using data from the referent registry generally bias the estimate of the parameter associated with family history away from the null even when risk of disease do not aggregate within families. This is sized-biased sampling; $\pi(\boldsymbol{y}_i) \propto n_i$.

In operational risk theory, only losses exceeding a certain threshold get recorded. A naive analysis based on the recorded loss data generally underestimates the frequency of losses and overestimates the severity of losses. This form of biased sampling is known as truncated sampling.

In epidemiology, the case-control design is arguably the most widely-used form of biased sampling for studying rare outcomes. Analyses based on this design have been well-studied by Cornfield (1951), Anderson (1972), Prentice and Pyke (1979), Breslow and Day (1980), Scott and Wild (1991, 1997, 2001, 2003), Manski and Lerman (1977), and Manski and McFadden (1981). The appeal of this design stems from the fact that while measures of absolute risk are not estimable, the odds ratio – a measure of relative risk – can be estimated efficiently by maximum likelihood as though the sample was obtained from a prospective design (Prentice and Pyke, 1979).

Biased sampling in the context of clustered binary data has recently attracted a great deal

of attention. Three sampling schemes stand out in the literature: proband designs, stratified sampling designs, and sampling designs based on the total number of diseased individuals in the cluster. The proband design, also known as the case-control family design, is an extension of the case-control design in the following way:

- Probands are obtained through the usual case-control design.
- Disease status and covariates are then obtained from relatives of the probands.
- The resulting sample consists of two types of families, those where the proband is diseased and those where the proband is disease-free.

Whittemore (1995), Zhao et al. (1998), and Wang and Hanfelt (2009) proposed statistical procedures that adjust for this study design by considering the distribution of the relatives' outcome conditional on the probands' outcome. Liang and Pulver (1996), Liang and Beaty (2000), Laird and Cuenco (2003), and Hudson et al. (2001) discussed logistic regression models where the non-proband marginal mean structure is modeled as a function of covariates and of the probands' outcomes. The parameter associated with the probands' outcome is interpreted as a measure of familial aggregation.

With stratified sampling, it is assumed that the population of clusters can be partitioned into a finite number of disjoint strata, denoted by $\mathcal{S}_1, \ldots, \mathcal{S}_J$. Stratum definition depends on the outcome vector of interest. Each cluster in the population belongs to exactly one stratum. There are two distinct versions of stratified sampling: standard stratified (SS) sampling and variable probability (VP) sampling. In SS sampling, a random sample of size N_j is obtained from the j-th stratum. In VP sampling, the study sample is obtained by repeating the following procedure:

1. randomly select a unit from the population and observe the pattern of the outcome vector
2. if the outcome vector belongs in $\mathcal{S}_j$, keep the unit with probability π_j

Lawless et al. (1999), Scott and Wild (2001), Neuhaus et al. (2002), and Neuhaus et al. (2006) developed semi-parametric maximum likelihood methods for stratified sampling of families.

The third form of biased sampling may be viewed as a special case of stratified sampling. Strata definition depends on the total of the response vector. This design is used frequently when the outcome of interest is rare. In this situation, researchers preferentially sample families where the total number of diseased members is at least some prespecified amount, say m. For example, if the total number of diseased members is at least m, then a family is selected to be in the sample with probability 1; otherwise, the family is selected with probability 0. Pfeiffer et al. (2001) examine the situation where $m = 2$ in the context of a GLMM with a random family effect and random genetic effects associated with each member. The case where $m = 1$ is considered by Burton et al. (2000), Glidden and Liang (2002), Epstein et al. (2002), Noh et al. (2005), and Neuhaus and Jewell (1990). Qaqish et al. (1997) examined this situation in the context of marginal models. They obtained explicit expressions for the sample mean and dependence structures.

3.4 Previous Works

3.4.1 Proband Design

Prentice and Pyke (1979) showed that data obtained from the case-control design can be viewed as a random sample from a pseudo distribution. By this we mean that the case-control sample can be thought of as a random sample from a population where the outcome distribution places a much greater mass on "success" than in the actual population to which you want to make inferences. Under a specified logistic regression model, they showed that the logistic regression model obtained from the pseudo model differs from the specified logistic regression model by an offset. Hence, they suggested that if the odds ratio is of interest, we could effectively ignore the sampling scheme and fit a logistic regression model as though the data was prospectively or randomly sampled. Since the proband design is an extension of the case-control design, it is natural to think that some extension of the results of Prentice and Pyke is possible.

Indeed, in the parametric setup, Whittemore (1995) showed that the proband design induces a retrospective likelihood with a retrospective component that is based entirely on the probands' data and a prospective component based on the relatives' data. The retrospective component

uses the result of Prentice and Pyke while the prospective component is derived by specifying a joint distribution for $\boldsymbol{Y}|\boldsymbol{X}$ based on the Bahadur (1961) representation. For constant family size of 2, this task is relatively straightforward. However, for large cluster size or variable cluster size, this task is not feasible.

Zhao et al. (1998) proposed two sets of estimating functions for case-control family studies. The first set models the relationship between each relative's outcome and their covariates while conditioning on the proband's outcome. This set of estimating functions utilizes information on pairwise correlations between a proband and his/her relative conditional on the proband's outcome. The second set of estimating functions utilizes information on pairwise correlations between two relatives, conditional on the proband's outcome. As with Whittemore's approach, this approach uses Bahadur's representation for specifying joint distributions. Unlike Whittemore's approach which uses Bahadur's representation to specify $\Pr(\boldsymbol{Y}_i|\boldsymbol{X}_i)$, this approach uses Bahadur's representation for specifying only two sets of distributions: $\Pr(Y_{ij}, Y_{i0}|\boldsymbol{X}_i)$ and $\Pr(Y_{ij_1}, Y_{ij_2}, Y_{i0}|\boldsymbol{X}_i)$ where j indexes the relatives outcome and Y_{i0} denotes the proband's outcome. This overcomes the cluster size limitation seen in Whittemore's approach.

Wang and Hanfelt (2009) considered the case-control family design in a finely stratified setup. This means that the marginal model adjusts for many stratum-specific effects by specifying stratum-specific intercepts in the model for the mean structure. Here, strata are treated as covariates and are not part of the sampling scheme. Under this finely stratified setup, they noticed that the proband-based estimating functions of Zhao et al. (1998) cannot be used due to the presence of many nuisance parameters. Building on their earlier works, they proposed adding correction terms to the proband-based estimating functions of Zhao et al. using their adjusted profile estimating function techniques (Wang and Hanfelt, 2008). There are three shortcomings of their approach. As noted in their paper, they are unable to make corrections on the proband component of their estimating functions. Second, a selling point of nuisance parameter insensitivity is based on asymptotics of the number of family within each stratum. This is not achievable in the finely stratified setup since the number of families in each stratum is few (Wang and Hanfelt, 2009, p.365, 3rd paragraph). Third, this approach applies to the situation where the outcome of interest is relatively common.

Some authors prefer to ignore the marginal information about the proband altogether. Instead they specify a population familial aggregation marginal model that treats the probands' outcome as a covariate:

$$\text{logit}\,\Pr(Y_{ij} = 1|y_{i0}, \boldsymbol{x}_{ij}) = \alpha_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_1 + \gamma y_{i0} \ . \tag{3.3}$$

See for example works by Liang and Pulver (1996) and Liang and Beaty (2000). Here, the interest is on γ, a measure of familial aggregation. We can interpret γ as follows: for fixed $\boldsymbol{x}$, e^γ is the odds ratio of disease between a relative and a case proband versus a control proband. However, if interest is on the effects of $\boldsymbol{x}$ on disease outcome without conditioning on Y_0, then model (3.3) cannot provide that answer.

We comment that proband-based sampling is, in general, not well-defined. As described in Whittemore (1995), the case families are obtained by sampling from the distribution $\Pr(\boldsymbol{Y}_{i-0}, \boldsymbol{Z}|Y_0 = 1)$ while control families are obtained by sampling from $\Pr(\boldsymbol{Y}_{i-0}, \boldsymbol{Z}|Y_0 = 0)$ where $\boldsymbol{Y}_{i-0} = (Y_{i1}, \ldots, Y_{in_i})^\top$. This is equivalent to saying that case families are obtained by sampling families from the stratum where all probands are diseased and measuring the outcomes of the probands' relatives as well as the covariates of everyone in the family. Similarly for control families. But the idea of stratifying the population of families based on proband status requires knowing who the proband is for every family in the population. It is not clear who gets to be the proband. The proband concept is a device for identifying members of a case-control sample and it is unclear how to extend this idea to families in the population.

3.4.2 Stratified Sampling

For logistic regression models, Anderson (1972) and Prentice and Pyke (1979) showed that under the case-control design, all parameters except the intercept are estimable. Furthermore, these parameters can be estimated by maximizing a likelihood that assumes the data was obtained prospectively. Inferences on the intercept, and hence absolute probabilities, cannot be made unless information on the sampling rates is available; in which case, it enters the regression model as offsets. For other link functions, information on the sampling rates and the

population distribution of the outcome variable is required to make inferences on any parameter in the specified regression model.

Scott and Wild (1997) generalized the results of Prentice and Pyke to include other link functions. They developed a maximum likelihood approach that utilizes extra information about population totals. For example, in a case-control study where the population has a known size N, their method requires that the total number of controls (N_0) in the population, the total number of cases in the population (N_1), and the sampling fractions (n_0/N_0 and n_1/N_1) are known. Lawless et al. (1999) developed semiparametric maximum likelihood methods that enable population parameters to be estimated under various biased sampling schemes that depend on both response and covariates; i.e., sampling is based on $\pi(Y, X)$.

These ideas are extended to the multivariate case by Neuhaus et al. (2002) to cover the analysis of retrospective family studies. The sampling scheme to which this extension applies is the stratified sampling scheme:

1. Stratify the population of families into L well-defined strata, denoted by $\mathcal{S}_1, \ldots, \mathcal{S}_L$.

2. Take a random sample of size m_ℓ from $\mathcal{S}_\ell$.

If we denote by $\boldsymbol{Y}_{\ell i}$ the response vector of the i-th family in ℓ-th stratum that was sampled and by $\mathcal{D}_\ell$ the set of families sampled from the ℓ-th stratum, then the above sampling scheme induces the following likelihood:

$$L(\boldsymbol{\theta}, g) = \prod_{\ell=1}^{L} \prod_{i \in \mathcal{D}_\ell} \left\{ \frac{f(\boldsymbol{y}_{\ell i} | \boldsymbol{X}_{\ell i}; \boldsymbol{\theta}) g(\boldsymbol{X}_{\ell i})}{\Pr(\boldsymbol{Y}_{\ell i}) \in \mathcal{S}_\ell} \right\} \tag{3.4}$$

The density function of $\boldsymbol{X}$, denoted by g, is not of interest and hence regarded as a nuisance parameter which is possibly infinite-dimensional depending on the support of $\boldsymbol{X}$. The case in which $\boldsymbol{X}$ has finite support and the outcome is univariate was addressed by Scott and Wild (1997) and Lawless et al. (1999). An alternative is to treat g nonparametrically and maximize the profile likelihood

$$\ell_P(\boldsymbol{\theta}) := \log L(\boldsymbol{\theta}, \hat{g}(\theta)) = \sum_{g} \log L(\boldsymbol{\theta}, g) \tag{3.5}$$

If g is infinite-dimensional, obtaining (3.5) is difficult.

To get around this, Neuhaus et al. (2002) considered a different sampling scheme where the sample is obtained as follows:

1. Randomly pick a family from the population of families and observe $(\boldsymbol{Y}, \boldsymbol{X})$
2. If $(\boldsymbol{Y}, \boldsymbol{X}) \in \mathcal{S}_\ell$, then keep $(\boldsymbol{Y}, \boldsymbol{X})$ with probability π_ℓ.
3. Repeat.

This is the VP sampling scheme described earlier (see Lawless et al., 1999):

$$\pi(\boldsymbol{y}_i, \boldsymbol{X}_i) = \pi_\ell \quad \text{if} \quad \boldsymbol{y}_i \in \mathcal{S}_\ell \ . \tag{3.6}$$

The induced sample model is

$$f(\boldsymbol{y}_i | \boldsymbol{X}_i; S_i = 1) = \frac{\pi(\boldsymbol{y}_i) f(\boldsymbol{y}_i | \boldsymbol{X}_i; \boldsymbol{\theta})}{\sum_{k=1}^{K} \sum_{\{\boldsymbol{y}: \boldsymbol{y} \in \mathcal{S}_k\}} \pi_k f(\boldsymbol{y} | \boldsymbol{X}_i; \boldsymbol{\theta})} \ . \tag{3.7}$$

By introducing stratum labels, we can rewrite (3.7) as

$$f(\boldsymbol{y}_{\ell i} | \boldsymbol{X}_{\ell i}; S_i = 1) = \frac{\pi_\ell f(\boldsymbol{y}_{\ell i} | \boldsymbol{X}_{\ell i}; \boldsymbol{\theta})}{\sum_{k=1}^{K} \sum_{\boldsymbol{y}: \boldsymbol{y} \in \mathcal{S}_k} \pi_k f(\boldsymbol{y} | \boldsymbol{X}_{\ell i}; \boldsymbol{\theta})}$$

which is the contribution to the likelihood from families from the ℓ-th stratum. The log likelihood takes the form

$$\ell(\boldsymbol{\theta}, \boldsymbol{\delta}) = \sum_{\ell=1}^{L} \sum_{i \in \mathcal{D}_\ell} \log f(\boldsymbol{y}_{\ell i} | \boldsymbol{X}_{\ell i}; S_i = 1) \tag{3.8}$$

where

$$\delta_\ell = \log\left(\frac{\pi_\ell}{\pi_L}\right), \quad \ell = 1, \ldots, L-1 \ .$$

The authors noted that the profile likelihood of (3.8) is equal to the profile likelihood given in (3.5):

$$\ell(\boldsymbol{\theta}, \hat{g}(\boldsymbol{\theta})) = \ell(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(\boldsymbol{\theta}))$$

Thus, even though $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(\boldsymbol{\theta}))$, is obtained from a sampling scheme that is different from the actual stratified sampling scheme, it is much easier The simplicity stems from the fact that $\boldsymbol{\delta}$

is finite-dimensional whereas g is infinite-dimensional. Furthermore, the estimation procedure for the finite-dimensional nuisance parameter was already developed by Lawless et al. (1999) and Scott and Wild (1997). An adjustment term is added to the information matrix to provide valid standard errors under the actual sampling scheme.

There are several difficulties with this approach foremost of which is the specification of the joint distribution of the response vector $f(\boldsymbol{y}_i|\boldsymbol{X}_i;\boldsymbol{\theta})$. For multivariate outcomes, this is not something that is easily done. The authors suggested in the introduction that $f(\boldsymbol{y}_i|\boldsymbol{X}_i;\boldsymbol{\theta})$ can be constructed using copulas (Meester and Mackay, 1994), or marginally specified mixed models (Heagerty, 1999). Even in the simple random sample situation, these models can be difficult to fit. It is not unreasonable to anticipate that this task is even more difficult in view of the biased sampling model (3.7). The alternative is to use either the Palmgren model or the Bahadur representation (which the authors used in their examples and simulations) but this is useful only for cluster size 2. A second difficulty is the estimation of the nuisance parameter $\boldsymbol{\delta}$. As the authors have noted, while these parameters are theoretically identifiable, they are practically not identifiable in the sense that $\hat{\boldsymbol{\delta}}$ is highly correlated with estimates of the intercept. The high correlation leads to convergence failure. Population stratum quantities and sampling fractions are needed in order to alleviate these problems (Neuhaus et al., 2002, see expression (5)). This information may or may not be available.

The generalized linear mixed model (GLMM) framework enables investigators to seek out parameters with cluster-specific interpretations. Although the biased sampling ideas developed by Neuhaus et al. (2002) apply to marginal models, it can accommodate GLMMs without change by viewing $f(\boldsymbol{y}_i|\boldsymbol{X}_i;\boldsymbol{\theta})$ in expression (3.7) as the result of integrating out the random effects. Neuhaus et al. (2006) applied this idea to GLMMs with logit link. They also proposed modifications based on the finite population framework to aid convergence caused by the fact that the variance components and the nuisance parameter $\boldsymbol{\delta}$ are all closely related to the intercept term.

3.4.3 Sampling Based on the Total

Neuhaus and Jewell (1990) discussed sampling where the sampling weights depend on the total

number of diseased outcomes in the cluster. For brevity, they refer to this form of retrospective sampling as PNR. In many respects, PNR may be viewed as a special case of stratified sampling in the sense that stratification of clusters depends on the total number of diseased individuals in the cluster. Their results can be summarized as follows:

- For a specified logistic regression model, PNR sampling generally leads to biased estimates of regression parameters if the sampling scheme is ignored. This statement applies to both marginal models and GLMMs.

- Under PNR sampling and for a specified random intercept GLMM with logit link, the method of conditional likelihood can be used to obtain consistent estimates of slope parameters associated with cluster-varying covariates while ignoring the sampling scheme.

- For covariates that have low variability, even conditional likelihood cannot be used to obtain parameters associated with these covariates.

- For outcome dependent sampling other than PNR, the sampling probabilities must be accounted for.

Qaqish et al. (1997) considered a specific form of PNR sampling whereby clusters are sampled according to

$$\pi(\boldsymbol{Y}) = \begin{cases} \pi_0 & \text{if } T = 0 \\ \pi_1 & \text{if } T \geq 1 \end{cases}$$

For a specified logistic regression model, they showed that the sampling scheme induces a sample marginal mean model that depends in a complicated way on the sampling ratio and the distribution of the total number of disease individuals in the cluster (c.f. Qaqish et al., 1997, expression 3). Similarly, the sample dependence structure (defined as pairwise odds ratio) also depends in a complicated way on the sampling ratio and the distribution of the total number of disease individuals in the cluster (c.f. Qaqish et al., 1997, expression 5).

When the sampling rates are known, Cai et al. (2001) suggested a weighted estimating

equations approach to parameter estimation for marginal models:

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{S_i}{\pi_i} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^{\top} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \ .$$

This idea of weighting by the inverse of the sampling rates goes back to Horvitz and Thompson (1952). A drawback of this approach is that it requires that $\min\{\pi_0, \pi_1\} > c \gg 0$. This rules out truncation; for example, if $\pi_0 = 0$ when $T_i = 0$, then this approach cannot be used.

Similar approaches have been suggested under parametric setups in which a weighted optimization function is maximized (see Wooldridge, 1999, 2001).

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{S_i}{\pi_i} \log \Pr(\boldsymbol{Y}_i | \boldsymbol{X}_i; \boldsymbol{\theta})$$

As with weighted GEE, this approach does not cover truncation. And as suggested earlier, specifying the joint distribution is difficult unless the cluster size is 2.

In genetic epidemiology, random sampling is rarely used. Instead, the intensity with which families are sampled depends on the number of affected members in the family with preferences given to families with at least one affected member. Depending on the type of study (linkage analysis, segregation analyses, etc.), various proposals have been suggested for dealing with ascertainment. As discussed in Thompson (1993), these various methods are generally invalid. This was illustrated by Burton et al. (2000). The probability model relating disease status with covariates that they considered was a random intercept GLMM with logit link. Families were obtained by complete ascertainment:

$$\pi(\boldsymbol{Y}) = \begin{cases} 0 & \text{if } T = 0 \\ 1 & \text{if } T > 0 \end{cases}$$

Using an ascertainment adjustment procedure based on Elston and Sobel (1979), they showed that the estimated variance components and regression parameters were not estimating the corresponding values in the population.

Glidden and Liang (2002) re-examined this issue. For a simple variance component model

with no covariates, they developed a likelihood that enabled them to obtain consistent estimates under complete ascertainment. However, their simulation study was limited in several ways. First, unlike the model considered by Burton et al., their model had no covariates. Second, their likelihood function assumes a constant family size (in their study, 5). Third, for small family sizes, the parameters of interest are not identifiable (Epstein, 2002). Fourth, their results require a correct specification of the mixing distribution. Under complete ascertainment, a misspecification of the random effect distribution can severely bias estimates (Glidden and Liang, 2002; Epstein, 2002).

Underscoring the need for methods that appropriately adjust for the sampling scheme, Epstein et al. (2002) noted that for the model and sampling scheme described by Burton et al. (2000), the correct way to adjust for the sampling scheme is to condition on the ascertainment event. That is, the likelihood to maximize should be based on $\Pr(\boldsymbol{Y}_i|\boldsymbol{X}_i, S_i = 1)$. They showed that the method described in Burton et al. inappropriately conditions on both the random effects and the ascertainment event: $\Pr(\boldsymbol{Y}_i|b_i, \boldsymbol{X}_i, S_i = 1)$. In some cases, even if an expression for $\Pr(\boldsymbol{Y}_i|\boldsymbol{X}_i, S_i = 1)$ is available, it may be too complex to be practical.

Pfeiffer et al. (2001) considered a GLMM with a random cluster effect and random within-cluster genetic effects under the scheme

$$\pi(\boldsymbol{Y}_i) = \begin{cases} 0 & \text{if } T_i < 2 \\ 1 & \text{if } T_i \geq 2 \end{cases}$$

They developed a likelihood that conditions on $T_i = 2$. There are two shortcomings of this approach. First, to adjust for the sampling scheme, they should condition on $T_i \geq 2$. Second, the data obtained from this sampling scheme provides very little information on the intercept and the variance component associated with the random cluster effect. This implies a very flat likelihood. To overcome numerical difficulties, they proposed a grid search. However, grid searches can be computationally intensive.

Bowden et al. (2007) proposed a two-stage data augmentation technique developed by Clayton (2003) that corrects for ascertainment bias. A simulation study suggests that their method

has some promise. In practice however, this method is also difficult to use because it requires generating extra data from a distribution whose parameter is unknown. In a simulation setting, it is easy to provide a parameter value close to the true value since we know *a-priori* what the true value is. No practical guidance is provided for choosing parameter values for generating the additional data.

3.5 Sampling Based on the Total: A Closer Look

Summary of Previous Works

In the previous section, we reviewed three sampling schemes associated with clustered data: proband sampling, stratified sampling, and sampling based on the total. We saw that sampling based on the total can be viewed as a special case of stratified sampling. We also saw that weighted approaches based on estimating equations and likelihoods are available for use if the sampling rates are known provided that the sampling rates are bounded away from zero. For marginal models, the likelihood approach has a major drawback in that it is very difficult to specify the joint distribution of the responses given the covariates. When the sampling rates are unknown, these weighted methods can still be used if consistent estimates of the weights are available. When the sampling rates are unknown, likelihood approaches based on the biased sampling distribution (c.f. Neuhaus et al., 2002) may be used. In addition to the difficulty of specifying the joint distribution of the responses given the covariates, this approach tends to produce likelihoods that are difficult to fit in the sense that optimization routines encounter convergence issues.

For mixed models, the likelihood induced by the sampling scheme are either flat or numerically difficult to maximize. Furthermore, misspecification of the mixing distribution can cause severe bias.

Outline of Goals

With the previously discussed background information in mind, we revisit sampling based on the total with a view towards developing an estimating function that does not require knowledge

of the sampling intensities and enable us to avoid some of the problems associated with the likelihood approach. Specifically, we

- derive the sample marginal model for a specified population marginal model
- construct an estimating function based on the sample marginal model
- derive the sample model for a specified population GLMM
- construct an estimating function based on the sample GLMM
- provide a simulation study to assess the performance of our estimating function in the marginal model setup
- provide a simulation study to assess the performance of our estimating function in the GLMM setup
- provide a non-numerical example that connects our method to a study design used in injury prevention research
- provide a numerical example that illustrates the use of our estimating function for analyzing twin pair data

3.6 Characterizing the Outcome-Dependent Sampling

Our development rests on the following characterization of the outcome-dependent sampling problem. Suppose there exist an infinite population of clusters (families, for example) where each cluster can be characterized by a response vector $\boldsymbol{Y}$. For the i-th cluster, we assume that $\boldsymbol{Y}$ is $n_i \times 1$ and write $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$. Suppose we can partition this population of clusters into R disjoint strata where the partitioning is based on some function of the response vector. Let $\mathcal{S}_r$ denote the r-th stratum. Sampling is done as follows: if $\boldsymbol{y}_i \in \mathcal{S}_r$, select the i-th cluster to be in the sample with probability π_r. If we denote the sampling intensity function by $\pi(\cdot)$, then this sampling scheme can be expressed as

$$\pi(\boldsymbol{y}_i) = \pi_r \quad \text{if} \quad \boldsymbol{y}_i \in \mathcal{S}_r \ . \tag{3.9}$$

Once a cluster is selected, we measure $\boldsymbol{Z}_i := (\boldsymbol{X}_i, \boldsymbol{W}_i)$ where $\boldsymbol{X}_i$ is an $n_i \times p$ matrix of cluster-varying covariates and $\boldsymbol{W}_i$ is an $n_i \times q$ matrix of cluster-constant covariates. The goal is to make inferences about the relationship between $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$ in the population.

The next section examines the situation where the relationship between $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$ in the population is described by a marginal regression model.

3.7 Marginal Models

Assume that the relationship between $\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$ is described by the following marginal regression model:

$$h(\mu_{ij}) = \beta_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_1 + \boldsymbol{w}_i^\top \boldsymbol{\beta}_2, \quad i = 1, \ldots, K, \quad j = 1, \ldots, n_i \tag{3.10}$$

where $\mu_{ij} = \Pr(Y_{ij} = 1; \boldsymbol{z}_{ij})$ is the marginal mean of Y_{ij} in the population and h is some link function. For prospectively or randomly sampled data, the parameters in (3.10) can be estimated by GEE (Liang and Zeger, 1986). In this section, we characterize the marginal mean of Y_{ij} induced by (3.9) and show why GEE is generally not applicable in this situation. When h is the log function, we show that (3.9) induces a sample model whose mean structure is similar to (3.10). We will show that this similarity in structure between the sample model and the population model enables us to prescribe an estimating equation for estimating $\boldsymbol{\beta}_1$.

Let S_i denote a sampling indicator where $S_i = 1$ means that cluster i was selected to be in the sample and where $S_i = 0$ means that cluster i was not selected to be in the sample. Let $\nu_{ij} = \Pr(Y_{ij} = 1 | S_i = 1; \boldsymbol{z}_{ij})$ denote the marginal mean in the sample. We will use the phrase "sample model" to mean $h(\nu_{ij})$ which is some function of $\boldsymbol{\beta} := [\beta_0,\ \boldsymbol{\beta}_1^\top,\ \boldsymbol{\beta}_2^\top]^\top$. In the presentations to follow, we obtain ν_{i1} for member "one" in the cluster but identical steps may be taken to obtain ν_{ij} for the j-th member. The quotes is meant to suggest that there is no notion of position. For notational convenience, define

$$\boldsymbol{y}_{-1} = (y_2, \ldots, y_{n_i}) \ , \quad \text{and}$$

$$p(y_1, \ldots, y_{n_i}; \boldsymbol{Z}_i) = \Pr(Y_{i1} = y_1, \ldots, Y_{in_i} = y_{n_i}; \boldsymbol{Z}_i) \ .$$

We have

$$\begin{aligned}\frac{\nu_{i1}}{1-\nu_{i1}} &= \frac{\Pr(Y_{i1}=1, S_i=1; \boldsymbol{Z}_i)}{\Pr(Y_{i1}=0, S_i=1; \boldsymbol{Z}_i)} \\ &= \frac{\sum_{\mathcal{A}_{i11}} \pi(\boldsymbol{y}_+) p(1, \boldsymbol{y}_{-1}; \boldsymbol{Z}_i) + \cdots + \sum_{\mathcal{A}_{i1R}} \pi(\boldsymbol{y}_+) p(1, \boldsymbol{y}_{-1}; \boldsymbol{Z}_i)}{\sum_{\mathcal{A}_{i01}} \pi(\boldsymbol{y}_-) p(0, \boldsymbol{y}_{-1}; \boldsymbol{Z}_i) + \cdots + \sum_{\mathcal{A}_{i0R}} \pi(\boldsymbol{y}_-) p(0, \boldsymbol{y}_{-1}; \boldsymbol{Z}_i)} \end{aligned} \tag{3.11}$$

where $\boldsymbol{y}_+ = (1, y_2, \ldots, y_{n_i})^\top$, $\boldsymbol{y}_- = (0, y_2, \ldots, y_{n_i})^\top$, and

$$\mathcal{A}_{i1r} = \{(y_2, \ldots, y_{n_i}) : \boldsymbol{y}_+ \in \mathcal{S}_r\}, \quad \mathcal{A}_{i0r} = \{(y_2, \ldots, y_{n_i}) : \boldsymbol{y}_- \in \mathcal{S}_r\} \ .$$

For example, if $\mathcal{S}_1$ defines a set of clusters whose response vector has a total of 1 event, then $(1, 0, \ldots, 0) \in \mathcal{S}_1$. In this example, $\mathcal{A}_{i11}$ contains the singleton $\mathcal{A}_{i11} = \{(1, 0, \ldots, 0)\}$ but $\mathcal{S}_1$ contains any vector where the total is 1 (there are n_i of them). On the other hand, $\mathcal{A}_{i01}$ is a set that has cardinality $n_i - 1$.

Note that in the summation given in (3.11), the index sets $\mathcal{A}_{iar}$ are allowed to have cardinality zero. For example, if $\mathcal{S}_2$ defines a stratum consisting of all clusters whose response vector has zero events, then the set $\mathcal{A}_{i12}$ is an empty set since there are no values of $(y_2, \ldots, y_{n_i})$ that will make the response vector $(1, y_2, \ldots, y_{n_i})$ have no events.

Sampling based on the total can be considered as special cases of (3.9). For example, if we stratify the population of clusters into two strata - one consisting of clusters that have no diseased member ($\mathcal{S}_1$) and one consisting of clusters that have at least one diseased member ($\mathcal{S}_2$), then (3.9) can be written as

$$\pi(\boldsymbol{y}_i) = \begin{cases} \pi_1 & \text{if } \boldsymbol{y}_i \in \mathcal{S}_1 \\ \pi_2 & \text{if } \boldsymbol{y}_i \in \mathcal{S}_2 \end{cases} = \begin{cases} \pi_1 & \text{if } t_i = 0 \\ \pi_2 & \text{if } t_i > 0 \end{cases}$$

This dissertation examines two cases of sampling based on the total: sampling clusters based on a pre-specified number of events and sampling clusters that exhibit variation in the outcome vector. For each sampling scheme, we derive exact expressions for the sample marginal model. Important features of the sample marginal model will be discussed in connection with the nuisance parameter problem. We propose an estimating equation for estimating the slope parameter when the population marginal mean model is described by the log link. In the subsections that follow we examine two special cases of (3.9) and their effects on (3.11).

3.7.1 Sampling Clusters Based on the Number of Events

The first sampling scheme under consideration samples clusters based on a pre-specified number of events in the cluster, which we denote by m. This can be expressed as

$$\pi(\boldsymbol{y}_i) = \begin{cases} \pi_1 & \text{if } t_i \leq m \\ \pi_2 & \text{if } t_i > m \end{cases} \tag{3.12}$$

where $t_i = \sum_j y_{ij}$. This design can be thought of as a generalization of the standard case-control study but cases and controls are not individuals but clusters. We can think of control clusters as clusters with no more than m diseased individuals ($t_i \leq m$) and case clusters as clusters with at least $m + 1$ diseased individual ($t_i > m$). In addition, this design defines two strata. Stratum $\mathcal{S}_1$ consists of all control clusters in the population. Stratum $\mathcal{S}_2$ consists of all case clusters. In practice, clusters with $t_i > m$ are preferentially sampled so that $\pi_2 \gg \pi_1$; for example, $\pi_1 = 0$.

A straightforward application of (3.11) shows that

$$\frac{\nu_{i1}}{1-\nu_{i1}} = \frac{\pi_1 \sum_{\mathcal{A}_{i11}} \Pr(Y_{i1} = 1, \boldsymbol{y}_{-1}; \boldsymbol{Z}_i) + \pi_2 \sum_{\mathcal{A}_{i12}} \Pr(Y_{i1} = 1, \boldsymbol{y}_{-1}; \boldsymbol{Z}_i)}{\pi_1 \sum_{\mathcal{A}_{i01}} \Pr(Y_{i1} = 0, \boldsymbol{y}_{-1}; \boldsymbol{Z}_i) + \pi_2 \sum_{\mathcal{A}_{i02}} \Pr(Y_{i1} = 0, \boldsymbol{y}_{-1}; \boldsymbol{Z}_i)} \tag{3.13}$$

where

$$\mathcal{A}_{i11} = \{(y_2, \ldots, y_{n_i}) : (1, \boldsymbol{y}_{-1}) \in \mathcal{S}_1 \}, \quad \mathcal{A}_{i12} = \{(y_2, \ldots, y_{n_i}) : (1, \boldsymbol{y}_{-1}) \in \mathcal{S}_2 \},$$
$$\mathcal{A}_{i01} = \{(y_2, \ldots, y_{n_i}) : (0, \boldsymbol{y}_{-1}) \in \mathcal{S}_1 \} \quad \mathcal{A}_{i02} = \{(y_2, \ldots, y_{n_i}) : (0, \boldsymbol{y}_{-1}) \in \mathcal{S}_2 \} .$$

Neuhaus and Jewell (1990) and Qaqish et al. (1997) considered the special case $m = 0$. When $m = 0$, expression (3.13) reduces to

$$\frac{\nu_{i1}}{1 - \nu_{i1}} = \frac{\mu_{i1}}{1 - \mu_{i1} + \kappa_i}$$

or, more generally,

$$\frac{\nu_{ij}}{1 - \nu_{ij}} = \frac{\mu_{ij}}{1 - \mu_{ij} + \kappa_i} . \tag{3.14}$$

where

$$\kappa_i = \Pr(T_i = 0; \boldsymbol{Z}_i)\,(r_{12} - 1) \qquad \text{and} \quad r_{12} = \frac{\pi_1}{\pi_2}.$$

Expression (3.14) is equivalent to

$$\nu_{ij} = \frac{\mu_{ij}}{1 + \kappa_i} .$$

Note that under random sampling $r_{12} = 1$ so that $\kappa_i = 0$ and $\nu_{ij} = \mu_{ij}$. When clusters with $t_i > 0$ are preferentially sampled ($\pi_2 \gg \pi_1$), we have $-1 < \kappa_i < 0$ and $\nu_{ij} > \mu_{ij}$. If h is the log function, then

$$\log \nu_{ij} = \lambda_i + \beta_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_1 + \boldsymbol{w}_i^\top \boldsymbol{\beta}_2 \tag{3.15}$$

where $\lambda_i = -\log(1 + \kappa_i)$. Expression (3.15) is the sample model induced by this sampling scheme. We see that under the assumed population model, this sampling scheme induces a sample model that has a similar structure as the population model – linearity in the parameters. However, the sampling scheme introduces cluster-specific parameters $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_K)^\top$. If λ_i is known, then GEE can be used to estimate $\boldsymbol{\beta}$, treating λ_i as offsets. But λ_i is known if and only if we know the sampling ratios and the distribution of T_i as a function of $\boldsymbol{Z}_i$. These

quantities may or may not be known. When λ_i is unknown, the sample model introduces K cluster-specific nuisance parameters. Specifically, the number of nuisance parameters grows with the sample size - a situation known as the Neyman-Scott problem (Neyman and Scott, 1948). Note that if we write (3.15) as

$$\log \nu_{ij} = \delta_i + \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}_1 \tag{3.16}$$

where $\delta_i = \lambda_i + \beta_0 + \boldsymbol{w}_i^{\top}\boldsymbol{\beta}_2$, then it has the same structure as (2.15). Using results from section (2.7), $\boldsymbol{\beta}_1$ can be estimated as the root of the estimating function given in (2.28). Robust standard errors of $\widehat{\boldsymbol{\beta}}_1$ can be approximated by (2.31).

Note that for logistic regression models under truncated sampling, O'Neill and Barry (1995) proposed an unbiased estimating function that is devoid of the notion of the sampling rates. However, their estimating function is obtained by differentiating a log likelihood that assumes the outcomes within the cluster are independent. This implies that the standard errors are not valid for the estimated regression parameters.

The case $m > 0$ does not simplify. In fact,

$$\frac{\nu_{ij}}{1-\nu_{ij}} = \frac{\mu_{ij} + (r_{12}-1)\Pr(T_i \leq m, Y_{ij} = 1; \boldsymbol{Z}_i)}{1-\mu_{ij} + (r_{12}-1)\Pr(T_i \leq m, Y_{ij} = 0; \boldsymbol{Z}_i)} \ . \tag{3.17}$$

Expression (3.17) is equivalent to

$$\nu_{ij} = \frac{\mu_{ij} + (r_{12}-1)\Pr(T_i \leq m, Y_{ij} = 1; \boldsymbol{Z}_i)}{1 + (r_{12}-1)\Pr(T_i \leq m; \boldsymbol{Z}_i)} \ . \tag{3.18}$$

Two things are evident from (3.17) and (3.18). First, regardless of whether the specified population regression model is based on the identity, log, or logit link, the sample model cannot be expressed as a linear function of $\boldsymbol{\beta}$. Second, there are a great deal more nuisance parameters than the case $m = 0$. These include the joint distribution of (T_i, Y_{ij}) and the sampling ratio r_{12}. With this many nuisance parameters, it is not clear how to go about estimating $\boldsymbol{\beta}$ (or any of its components).

3.7.2 Sampling Clusters Exhibiting Variation

This sampling scheme depends on the observed variability of the response vector. An outcome vector exhibits variability if $\boldsymbol{Y}_i \neq \mathbf{0}, \mathbf{1}$. Note that if $\boldsymbol{Y}_i = \mathbf{0}$ then $t_i = 0$ and if $\boldsymbol{Y}_i = \mathbf{1}$ then $t_i = n_i$. From this observation, we can describe this design by

$$\pi(\boldsymbol{y}_i) = \begin{cases} \pi_1 & \text{if } t_i = 0 \text{ or } t_i = n_i \\ \pi_2 & \text{if } 0 < t_i < n_i \end{cases} \tag{3.19}$$

In practice it may be advantageous to preferentially sample clusters exhibiting variation so that $\pi_2 \gg \pi_1$. For example, suppose a prospective or longitudinal study has been conducted and certain bio-specimens (i.e.; blood samples, tissues) are collected on each individual. At some point in the future, researchers may want to investigate the relationship between the outcome and certain covariates such as genetic risk factors that can be measured from the bio-specimens through certain laboratory procedures. In cases where laboratory costs are expensive, it may not be feasible to obtain laboratory measurements for all bio-specimens. Since the informative clusters are those that exhibit variation in the response, it is suggested that only those clusters should be studied. Schildcrout and Heagerty (2008) discussed the merits of this sampling scheme in a parametric setup. Specifically, they suggested the following selection probabilities: $\pi_1 = 0$ and $\pi_2 = 1$.

As with the design defined in section 3.7.1, this design defines two strata. Stratum $\mathcal{S}_1$ consists of all clusters where either no member has any event or every member has the event. Stratum $\mathcal{S}_2$ consists of all clusters where the total number of events is at least 1 but less than the cluster size. Let

$$\begin{aligned} \mathcal{A}_{i11} &= \{(y_2, \ldots, y_{n_i}) : (1, \boldsymbol{y}_{-1}) \in \mathcal{S}_1\} , \\ \mathcal{A}_{i12} &= \{(y_2, \ldots, y_{n_i}) : (1, \boldsymbol{y}_{-1}) \in \mathcal{S}_2\} , \\ \mathcal{A}_{i01} &= \{(y_2, \ldots, y_{n_i}) : (0, \boldsymbol{y}_{-1}) \in \mathcal{S}_1\} , \\ \mathcal{A}_{i02} &= \{(y_2, \ldots, y_{n_i}) : (0, \boldsymbol{y}_{-1}) \in \mathcal{S}_2\} . \end{aligned}$$

An application of (3.11) gives

$$\begin{aligned}\frac{\nu_{i1}}{1-\nu_{i1}} &= \frac{\Pr(Y_{i1}=1, S_i=1; \boldsymbol{Z}_i)}{\Pr(Y_{i1}=0, S_i=1; \boldsymbol{Z}_i)} \\ &= \frac{\mu_{i1} + (r_{12}-1)\Pr(T_i = n_i; \boldsymbol{Z}_i)}{1-\mu_{i1} + (r_{12}-1)\Pr(T_i = 0; \boldsymbol{Z}_i)} \ .\end{aligned}$$

For the j-th member we have

$$\frac{\nu_{ij}}{1-\nu_{ij}} = \frac{\mu_{ij} + (r_{12}-1)\Pr(T_i = n_i; \boldsymbol{Z}_i)}{1-\mu_{ij} + (r_{12}-1)\Pr(T_i = 0; \boldsymbol{Z}_i)} \quad \text{where} \quad r_{12} = \frac{\pi_1}{\pi_2} \ . \tag{3.20}$$

Expression (3.20) is equivalent to

$$\nu_{ij} = \frac{\mu_{ij} + (r_{12}-1)\Pr(T_i = n_i; \boldsymbol{Z}_i)}{1 + (r_{12}-1)\big[1 - \Pr(0 < T_i < n_i; \boldsymbol{Z}_i)\big]} \ . \tag{3.21}$$

It can be seen from (3.21) that whether h is the identity, log, or logit link, $h(\nu_{ij})$, as a function of $\boldsymbol{\beta}$, does not have the same structure as the specified $h(\mu_{ij})$. If the disease is rare, then $\Pr(T_i = n_i; \boldsymbol{Z}_i) \approx 0$ so that

$$\nu_{ij} \approx \frac{\mu_{ij}}{1 + (r_{12}-1)\big[1 - \Pr(0 < T_i < n_i; \boldsymbol{Z}_i)\big]} \ . \tag{3.22}$$

In this case, if h is the log link, then

$$\log \nu_{ij} \approx \lambda_i + \beta_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_1 + \boldsymbol{w}_i^\top \boldsymbol{\beta}_2 \tag{3.23}$$

where $\lambda_i = -\log\big(1 + (r_{12}-1)\big[1 - \Pr(0 < T_i < n_i; \boldsymbol{Z}_i)\big]\big)$. This has approximately the same structure as (3.15) which enables $\boldsymbol{\beta}_1$ to be estimated by (2.28). Robust standard errors of $\widehat{\boldsymbol{\beta}}_1$ can be approximated by (2.31).

In some practical settings, the rare disease assumption does not have to hold strictly if n_i is large. Unless the disease is extremely common, it is highly unlikely that everyone in the cluster has the disease. Thus, it is reasonable to assume that $\Pr(T_i = n_i; \boldsymbol{X}_i) \approx 0$.

3.7.3 Summary of Outcome-Dependent Sampling For Marginal Models

In this section, we examine two special cases of biased sampling based on the total:

1. selecting clusters using weights that are based on the number of diseased members in the cluster

2. selecting clusters using weights that are based on discordance in the response vector

In item 1, we showed that when h is the log function, we can consistently estimate $\boldsymbol{\beta}_1$ – the parameter vector associated with cluster-varying covariates – if selection is based on partitioning the population of clusters into those that have no diseased member and those that have at least 1 diseased member. In item 2, we also showed that if h is the log function, we can obtain approximately consistent estimates of $\boldsymbol{\beta}_1$ if selection is based on partitioning the population of clusters into those whose response vector exhibit variation and those whose response vector do not exhibit variation. Estimation is done by utilizing (2.28) – an estimating function that confers several advantages. First, it is relatively simple to use. As discussed in section 2.9, $\boldsymbol{\beta}_1$ can be estimated by using existing software that fits the Cox proportional hazards model with the caveat that the standard errors from these software are invalid. Second, it does not require us to know the sampling ratios. This is convenient because in some situations, we may not have information on these ratios. And finally, no nuisance parameter has to be estimated. This is true when we solve the estimating function and when we compute the sandwich estimator.

3.8 Connections to the Double-Pair Design

In the field of injury prevention research, Evans (1986a,b) studied seat-belt effectiveness based on a study design which he called the double pair design – a variant of the Mantel-Haenszel method (Cummings et al., 2003). In this study design, two sets of two-occupant vehicles involved in fatal road traffic accidents are sampled. The aim is to estimate relative risk of death between belted and unbelted vehicle occupants. The design can be described by Table 3.1. Subtable 1 pertains to belted drivers and unbelted passengers while subtable 2 pertains to unbelted drivers and unbelted passengers. In subtable 1, there are a vehicles in the sample where the driver died

Table 3.1: Evans' double-pair design

	Passenger (Y_2)					
	subtable 1			subtable 2		
Driver (Y_1)	Survived (0)	Died (1)	**Total**	Survived (0)	Died (1)	**Total**
Survived (0)	0	b		0	k	
Died (1)	a	c	d	j	l	m
Total		e			n	

but the passenger survived. Similarly, there are b vehicles where the driver survived but the passenger died and there are c vehicles where both occupants died. Note, there are no vehicles in the sample where both occupants survived. In a collision, Evans suggested that the relative risk of death between a belted and an unbelted occupant,

$$\frac{\Pr(Y=1|X=1)}{\Pr(Y=1|X=0)}\,, \tag{3.24}$$

can be estimated by nd/me, where $Y=1$ means the occupant died and $X=1$ means the occupant wears a seatbelt. We will refer to this quantity as the Evans' estimator.

On closer examination, the double pair design can be cast as a special case of (3.12) where $m=0$ and $\pi_1=0$. As is clear from Table 3.1, clusters with no fatalities are not sampled ($\pi_1=0$). Thus the sampling scheme may be described by

$$\pi(\boldsymbol{y}_i)=\begin{cases}0 & \text{if } t_i=0\\ \pi_2 & \text{if } t_i>0\end{cases} \tag{3.25}$$

We will show that the Evans' estimator is a solution to (2.28). To see this, each vehicle involved in an accident is viewed as a cluster and the occupants (driver and passenger) as cluster members. Since Evans was interested in estimating the population relative risk of death

between belted and unbelted drivers, a reasonable population model is

$$\log \mu_{ij} = \alpha + d_{ij}\theta + s_{ij}\gamma + \boldsymbol{z}_i^\top \boldsymbol{\lambda} \; ; \quad i = 1, \ldots, K; \quad j = 1, 2 \tag{3.26}$$

where s_{ij} is a seatbelt usage indicator (1 for belted and 0 for unbelted), d_{ij} is a driver indicator (1 for driver and 0 for passenger), and $\boldsymbol{z}_i$ are unspecified car-level characteristics. The index i indexes the cars/crashes while the index j indexes the occupants. We will let $j = 1$ denote drivers and $j = 2$ denote passengers.

The parameters θ and γ have the following interpretation. For a given seatbelt status and fixed $\boldsymbol{z}_i$, the log relative risk of death between drivers and passengers is θ. For a given occupant status (say driver) and fixed $\boldsymbol{z}_i$, the log relative risk of death between drivers who wear seatbelts and those who do not is γ. The sampling scheme (3.25) induces the following sample model:

$$\log \nu_{ij} = \delta_i + d_{ij}\theta + s_{ij}\gamma \tag{3.27}$$

where $\delta_i = \alpha + \log\big[1 - \Pr(T_i = 0; \boldsymbol{X}_i)\big] + \boldsymbol{z}_i^\top \boldsymbol{\lambda}$. If we apply (2.28) to (3.27), then it can be shown that

$$e^{\hat{\gamma}} = \frac{nd}{me} \; ,$$

which is the Evans's estimator. The details are provided appendix A.1. This suggests that for cluster size 2, our approach recovers the Evans' estimator as a special case.

One of the main criticisms of the double pair design is that it does not adjust for variables related to the vehicle or the crash (Cummings et al., 2003). These include make, weight, speed, and rollover. As shown by (3.26) these characteristics can be represented by $\boldsymbol{z}_i^\top \boldsymbol{\lambda}$. The solution to our estimating equation suggests that Evans' estimator does adjust for these confounders.

A shortcoming of the Evans' double-pair design is that his estimator cannot adjust for continuous covariates such as age. To adjust for age, Evans (1986b) categorized age, compute the risk ratio for each age group, take a weighted average of the log of the computed risk ratios, and obtain the average of the risk ratios by exponentiating the weighted average of the log of the computed risk ratios. As discussed by Evans (1986a, pg 224, 2nd paragraph), this approach

can be biased. Furthermore, the justification for the chosen weights are ambiguous. On the other hand, if we use a modeling approach, as is done in (3.27), our estimating function enables us to generalize the Evans procedure to adjust for any continuous covariates with the added advantage that any number of passengers can be incorporated simultaneously. The double pair design can only incorporate a single passenger.

Implicit in our approach is the assumption of no interference – the covariate status of one member of the cluster does not affect the outcome of the other members. In the context of vehicular accidents, this may be unreasonable since an unbelted person may be projected in such a way during an accident as to cause the death of other occupants. It isn't clear whether our procedure still produces consistent estimates in this setting. However, it has been suggested by Pepe and Anderson (1994) that when the no interference assumption does not hold, then using GEE with independence working correlation may still produce unbiased estimates. Since our estimating equation was obtained by modifying a GEE with independence working correlation, it might be possible that it provides consistent estimates when there is interference. This is, at present, a conjecture which remains to be investigated in the future.

3.9 Simulations: Marginal Models and Outcome-Dependent Sampling

In this section, results of a simulation study to assess the performance of our method is presented. We consider correlated binary outcomes with cluster size 5 and assume that the marginal mean in the population is related to a covariate x in the following way:

$$h(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$$

where $\mu_{ij} = \Pr(Y_{ij} = 1; x_{ij})$ with $X \sim \text{Bernoulli}(p)$. Two link functions are considered: log and logit. Although our procedure is developed under the log link, under rare disease the risk ratio and the odds ratio are approximately equal. Hence, our interest in the logit link. Response vectors are generated using the conditional linear family method (Qaqish, 2003) with exchangeable correlation $\rho = 0.10$. We assume the following outcome-dependent sampling

scheme:

$$\pi(\boldsymbol{y}_i) = \begin{cases} 0 & \text{if } t_i = 0, \\ 1 & \text{if } t_i > 0 \end{cases}$$

Selection stops when there are K clusters in the sample with at least one diseased member. This is analogous to sampling N_K clusters until we obtain K clusters with at least 1 diseased member and using only these K clusters in the analysis.

3.9.1 Log Link

For the log link, we assume that the population model that relates the outcome to the covariate is described by

$$\log \mu_{ij} = \beta_0 + \beta_1 x_{ij}$$

It is assumed that $\beta_0 = -2.5$ and $\beta_1 \in \{0.1, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. Two covariate distributions are considered: $p = 0.2$ and $p = 0.5$. Prevalence is roughly 11% when $\beta_1 = 1$ and $p = 0.2$. At $\beta_1 = 1$ and $p = 0.5$, prevalence is roughly 15%. We sample until $K = 100$ clusters with $t_i > 0$ is obtained.

Table 3.2 shows that our estimating function performs exceptionally well with respect to consistency. Confidence intervals were constructed using the robust standard errors. Using this procedure for constructing 95% confidence intervals, we see that the coverage is close to the nominal value. It isn't clear from this simulation study whether our robust standard errors are consistently bigger or consistently smaller than the actual standard errors. Our data suggests that sometimes they are bigger and sometimes they are smaller. On average however, they are close to the actual standard errors.

Table 3.2 analyzes only the $K = 100$ clusters that enjoy $t_i > 0$. N_K tells us how many clusters, on average, that we needed to sample in order to obtain these 100 clusters. In the first row, we see that, on average, $N_K = 257$ clusters are needed. If instead of analyzing only the 100 clusters with $t_i > 0$, we analyze all clusters that were sampled to obtain these 100 clusters, we can get some idea about efficiency. Table 3.3 presents estimates and standard errors from

Table 3.2: 2000 simulations were performed. N_K is the average number of clusters that we needed to sample in order to obtain 100 clusters that have at least 1 diseased member. Estimation is done using the estimating function (2.28)

β_1	p	N_K	$\hat{\beta}_1$	True[a] SE($\hat{\beta}_1$)	Robust SE($\hat{\beta}_1$)	Coverage[b]
1.0	0.2	257	0.998	0.1792	0.1816	0.954
0.9	0.2	267	0.891	0.1872	0.1824	0.946
0.8	0.2	275	0.791	0.1881	0.1845	0.945
0.7	0.2	284	0.695	0.1881	0.1869	0.948
0.6	0.2	292	0.593	0.1928	0.1898	0.948
0.5	0.2	300	0.495	0.1921	0.1940	0.955
0.1	0.2	328	0.099	0.2196	0.2122	0.948
1.0	0.5	200	1.000	0.1903	0.1825	0.943
0.9	0.5	211	0.907	0.1806	0.1809	0.949
0.8	0.5	223	0.809	0.1829	0.1787	0.946
0.7	0.5	235	0.707	0.1788	0.1768	0.944
0.6	0.5	250	0.604	0.1751	0.1748	0.954
0.5	0.5	263	0.507	0.1755	0.1736	0.944
0.1	0.5	319	0.101	0.1747	0.1727	0.943

a: var($\hat{\beta}_1$)
b: 95% c.i. constructed using robust standard errors

these calculations. If we compare the true standard error in row 1 of table 3.3 with that from table 3.2, we see that the relative efficiency between using the negative binomial sample (which is approximately the same as a random sample of size 257) with the bias sample of size 100 is

$$\left(\frac{0.1545}{0.1792}\right)^2 = 0.7433$$

This suggests that, on average, we need a random sample of approximately 192 (0.7433×257) to achieve the same efficiency as a bias sample of 100 clusters with $t_i > 0$. Table 3.4 provides some ideas about the efficiency gains.

Table 3.3: Parameters are estimated using all sampled clusters, not just the 100 clusters with $t_i > 0$. Estimation procedure is GEE with exchangeable working correlation

β_1	p	N_K	$\hat{\beta}_1$	True SE($\hat{\beta}_1$)	Robust SE($\hat{\beta}_1$)	Coverage
1.0	0.2	257	0.995	0.1545	0.1564	0.956
0.9	0.2	267	0.893	0.1613	0.1585	0.946
0.8	0.2	275	0.793	0.1663	0.1618	0.949
0.7	0.2	284	0.692	0.1668	0.1651	0.952
0.6	0.2	292	0.589	0.1661	0.1689	0.957
0.5	0.2	300	0.492	0.1741	0.1736	0.958
0.1	0.2	328	0.095	0.2016	0.1957	0.948
1.0	0.5	200	1.010	0.1788	0.1720	0.945
0.9	0.5	211	0.909	0.1699	0.1699	0.946
0.8	0.5	223	0.808	0.1729	0.1673	0.947
0.7	0.5	235	0.710	0.1664	0.1651	0.954
0.6	0.5	250	0.603	0.1604	0.1631	0.951
0.5	0.5	263	0.508	0.1646	0.1617	0.944
0.1	0.5	319	0.101	0.1624	0.1604	0.944

Table 3.4: Efficiency calculations. K_{SRS} is the approximate sample size under random sampling that provides the same amount of efficiency as our biased sample of $K = 100$

	$p = 0.2$		$p = 0.5$	
β	N_K	K_{SRS}	N_K	K_{SRS}
1.0	257	192	200	177
0.9	267	199	211	187
0.8	275	215	223	200
0.7	284	224	235	204
0.6	292	217	250	210
0.5	300	247	263	232
0.1	328	277	319	276

Table 3.5 shows what happens when we ignore the sampling scheme. Estimates are inconsistent and the coverage is substantially below the nominal value. However, the simulation results suggest that when $\beta_1 = 0$, ignoring the sampling scheme still provides consistent estimates of this null value (not shown). This is not a generalization but rather an observation that pertains only to this simulation study.

Table 3.5: Estimation is done using GEE with log link and exchangeable working correlation. The sampling scheme is ignored

β_1	$\Pr(X=1)$	N_K	$\hat{\beta}_1$	True se($\hat{\beta}_1$)	Robust se($\hat{\beta}_1$)	Coverage
1.0	0.2	257	0.697	0.1341	0.1308	0.356
0.9	0.2	267	0.623	0.1379	0.1320	0.447
0.8	0.2	275	0.546	0.1411	0.1347	0.524
0.7	0.2	284	0.475	0.1414	0.1375	0.637
0.6	0.2	292	0.401	0.1407	0.1404	0.721
0.5	0.2	300	0.328	0.1478	0.1441	0.795
0.1	0.2	328	0.062	0.1669	0.1628	0.942
1.0	0.5	200	0.784	0.1583	0.1522	0.667
0.9	0.5	211	0.694	0.1502	0.1487	0.701
0.8	0.5	223	0.601	0.1518	0.1453	0.691
0.7	0.5	235	0.521	0.1463	0.1426	0.731
0.6	0.5	250	0.433	0.1422	0.1397	0.758
0.5	0.5	263	0.361	0.1430	0.1370	0.793
0.1	0.5	319	0.067	0.1367	0.1337	0.937

3.9.2 Logit Link

While our development depends on the assumption of a log link function (multiplicative risk model), it is well-known that when the outcome is rare, the odds ratio and the means ratio are approximately equal. This implies that, under the rare disease assumption, the slope parameter from a logistic regression model can be approximated using our procedure as though the true model was a multiplicative risk model. For the logit link, we assume that the relationship between the outcome and covariate is described by

$$\text{logit}\,\mu_{ij} = \beta_0 + \beta_1 x_{ij} \ .$$

Parameter values under consideration were $\beta_0 = -4.0$, $\beta_1 \in \{0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6\}$, and it is assumed that $X \sim \text{Bernoulli}(p = 0.5)$. Under the assumed covariate distribution, $\beta = 1.6$ translates to, roughly, a 5% prevalence. We sample until $K = 100$ clusters with $t_i > 0$ is obtained.

Table 3.6 provides an assessment of how well our method performs. First, we see that our

estimating function does a reasonable job of estimating the odds ratio. Even though there is some bias, the degree of the bias is very small. Second, even at $\beta_1 = 1.6$ (which corresponds to, roughly, 5.0% prevalence), our estimating function produces an estimate that is relatively close to 1.6. This suggests that we may still obtain reasonable estimates even up to $\beta_1 = 2.0$, which corresponds to roughly a 6.8% prevalence. Thus, our estimating function could limit the definition of rare disease to about 7 percent prevalence beyond which estimates of the odds ratio based on our method may be unreliable. Third, if we use the robust standard errors to construct 95% confidence intervals, the coverage is slightly below the nominal value.

Table 3.6 analyzes only the $K = 100$ clusters that enjoy $t_i > 0$. N_K tells us how many clusters, on average, that we needed to sample in order to obtain these 100 clusters. In the first row, we see that, on average, $N_K = 1100$ clusters are needed.

If instead of analyzing only the 100 clusters with $t_i > 0$, we analyze all 1100 clusters, we can obtain some idea about efficiency. Table 3.7 presents estimates and standard errors from these calculations. If we compare the true standard error in row 1 of table 3.7 with that from table 3.6, we see that the relative efficiency between using the negative binomial sample (which is approximately the same as a random sample of size 1100) with the bias sample of size 100 is 0.8814 $(0.1827/0.1946)^2$. This suggests that, on average, we need a random sample of approximately 970 (0.8814×1100) clusters to achieve the same efficiency as a bias sample of 100 clusters with $t_i > 0$. This is a substantial gain. Table 3.8 provides some ideas about the efficiency gains under the rare disease assumption. It can be seen that the rarer the outcome, the greater the efficiency gains. Table 3.9 shows that there is substantial bias if we analyze the 100 clusters with $t_i > 0$ ignoring the sampling scheme.

3.10 Generalized Linear Mixed Models (GLMMs)

In section 3.7, we discussed two biased sampling schemes based on the total in the context of marginal models. We show that not only can we estimate the slope parameter consistently

Table 3.6: 2000 simulations were performed. β_1 is estimated by (2.28). Robust standard error is estimated by (2.31)

β_1	p	N_K	$\hat{\beta}_1$	True SE($\hat{\beta}_1$)	Robust SE($\hat{\beta}_1$)	Coverage
0.4	0.5	1100	0.388	0.1946	0.1935	0.942
0.6	0.5	984	0.598	0.1990	0.1970	0.944
0.8	0.5	866	0.780	0.2082	0.2030	0.949
1.0	0.5	762	0.984	0.2145	0.2106	0.942
1.2	0.5	665	1.175	0.2255	0.2218	0.941
1.4	0.5	580	1.361	0.2400	0.2326	0.936
1.6	0.5	502	1.553	0.2449	0.2455	0.943

Table 3.7: β_1 is estimated using all N_K clusters, not just those with $t_i > 0$. Estimation procedure is GEE with logit link and exchangeable working correlation.

β_1	p	N_K	$\hat{\beta}_1$	True SE($\hat{\beta}_1$)	Robust SE($\hat{\beta}_1$)	Coverage
0.4	0.5	1100	0.398	0.1827	0.1840	0.951
0.6	0.5	984	0.611	0.1917	0.1881	0.946
0.8	0.5	866	0.799	0.1964	0.1949	0.957
1.0	0.5	762	1.011	0.2062	0.2034	0.948
1.2	0.5	665	1.217	0.2189	0.2149	0.948
1.4	0.5	580	1.417	0.2352	0.2265	0.945
1.6	0.5	502	1.622	0.2397	0.2401	0.958

Table 3.8: K_{SRS} is the approximate sample size under random sampling that provides the same amount of efficiency as our biased sample of $K = 100$

α_1	N_K	K_{SRS}
0.4	1100	970
0.6	984	914
0.8	866	771
1.0	762	705
1.2	665	627
1.4	580	558
1.6	502	481

Table 3.9: Estimation is done using GEE with logit link and exchangeable working correlation. The sampling scheme is ignored.

β_1	p	N_K	$\hat{\beta}_1$	True SE($\hat{\beta}_1$)	Robust SE($\hat{\beta}_1$)	Coverage
0.4	0.5	1100	0.269	0.1826	0.1799	0.875
0.6	0.5	984	0.432	0.1947	0.1834	0.815
0.8	0.5	866	0.565	0.1995	0.1873	0.733
1.0	0.5	762	0.730	0.2122	0.1934	0.668
1.2	0.5	665	0.905	0.2288	0.2009	0.647
1.4	0.5	580	1.064	0.2462	0.2064	0.598
1.6	0.5	502	1.252	0.2456	0.2155	0.604

using (2.28) when the link function is the log link, but that there is also a substantial gain in efficiency in using a biased sample rather than a random sample. In this section, we discuss biased sampling in the context of GLMMs. To this end, recall from section 3.2 that a GLMM is specified by

$$h(\mu_{ij}^c) = \alpha_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_1 + \boldsymbol{w}_i^\top \boldsymbol{\alpha}_2 + \boldsymbol{u}_{ij}^\top \boldsymbol{b}_i \quad i = 1, \ldots, K;\ j = 1, \ldots, n_i \tag{3.28}$$

where h is a link function and

- $\boldsymbol{b}_i$ is a $r \times 1$ random effect vector associated with the i-th cluster such that $\boldsymbol{b}_i \sim G(\boldsymbol{\theta})$ for some m-dimensional parameter $\boldsymbol{\theta}$
- $\text{cov}(\boldsymbol{b}_i, \boldsymbol{Z}_i) = \boldsymbol{0}$, where $\boldsymbol{Z}_i = [\boldsymbol{X}_i,\ \boldsymbol{W}_i]$
- $\mu_{ij}^c = E[Y_{ij}|\ \boldsymbol{b}_i, \boldsymbol{Z}_i]$
- $\boldsymbol{u}_{ij}$ is $r \times 1$ vector of covariates

Under prospective or random sampling, $\boldsymbol{\alpha} := (\alpha_0,\ \boldsymbol{\alpha}_1^\top,\ \boldsymbol{\alpha}_2^\top)^\top$ and $\boldsymbol{\theta}$ can be estimated by maximizing the following marginal likelihood:

$$L(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{i=1}^{K} \left\{ \int \prod_{j=1}^{n_i} \Pr(Y_{ij} = y_{ij} | \boldsymbol{b}, \boldsymbol{x}_{ij}, \boldsymbol{w}_i, \boldsymbol{z}_{ij}) dG(\boldsymbol{b}; \boldsymbol{\theta}) \right\} \tag{3.29}$$

If we apply (3.9) to clustered binary data, the biased sampling distribution associated with the i-th cluster is

$$\begin{aligned}
&\Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i | S_i = 1, \boldsymbol{X}_i, \boldsymbol{W}_i) \\
&\qquad = \frac{\pi(\boldsymbol{y}_i) \int \Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{b}, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{U}_i) dG(\boldsymbol{b}; \boldsymbol{\theta})}{\sum_{\ell=1}^{R} \pi_\ell \sum_{\tilde{\boldsymbol{y}} \in \mathcal{S}_\ell} \int \Pr(\boldsymbol{Y}_i = \tilde{\boldsymbol{y}} | \boldsymbol{b}, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{U}_i) dG(\boldsymbol{b}; \boldsymbol{\theta})} \ . \qquad (3.30)
\end{aligned}$$

This expression can be motivated as follows. For simplicity, assume Y is continuous and univariate. Assume also that b is univariate; i.e., $u_{ij} = 1$. Define 2 strata $\mathcal{S}_1$ and $\mathcal{S}_2$ by $\mathcal{S}_1 = \{y : y \leq 2\}$ and $\mathcal{S}_2 = \{y : y > 2\}$. Suppose the study design is

$$\pi(Y_i) = \begin{cases} \pi_1 & \text{if } Y_i \leq 2 \\ \pi_2 & \text{if } Y_i > 2 \end{cases}$$

An application of Bayes' theorem produces

$$\begin{aligned}
f(y_i | S_i = 1, \boldsymbol{x}_i, \boldsymbol{w}_i) &= \frac{f(S_i = 1, y_i, \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i)}{\int_{-\infty}^{\infty} p(S_i = 1, \tilde{y}, \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i) d\tilde{y}} \\
&= \frac{\Pr(S_i = 1 | y_i, \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i) f(y_i | \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i)}{\int_{-\infty}^{\infty} \Pr(S_i = 1 | \tilde{y}, \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i) f(\tilde{y} | \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i) d\tilde{y}} \\
&= \frac{\pi(y_i) f(y_i | \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i)}{\int_{-\infty}^{2} \pi_1 f(\tilde{y} | \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i) d\tilde{y} + \int_{2}^{\infty} \pi_2 f(\tilde{y} | \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i) d\tilde{y}} \\
&= \frac{\pi(y_i) \int f(y_i | b, \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i) dG(b; \boldsymbol{\theta})}{\sum_{\ell=1}^{2} \pi_\ell \int_{\tilde{y} \in \mathcal{S}_\ell} \int f(\tilde{y} | b, \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i) dG(b; \boldsymbol{\theta}) d\tilde{y}}
\end{aligned}$$

where $f(S_i = 1, y_i, \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{u}_i)$ is the joint distribution of Y_i, $\boldsymbol{X}_i$, $\boldsymbol{W}_i$, and $\boldsymbol{U}_i$ in the sample. For binary outcomes, the expression $\int_{\tilde{y} \in \mathcal{S}_\ell} (\cdot) d\tilde{y}$ is replaced by $\sum_{\tilde{y} \in \mathcal{S}_\ell} (\cdot)$ and the probability density function $f(y_i | \cdots)$ is replaced by $\Pr(Y_i = y_i | \cdots)$. This gives (3.30).

Theoretically, if the sampling rates are known, we can estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ by maximizing the following likelihood:

$$L(\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{i=1}^{K} \left\{ \frac{\pi(\boldsymbol{y}_i) \int \Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{b}, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{U}_i) dG(\boldsymbol{b}; \boldsymbol{\theta})}{\sum_{\ell=1}^{R} \pi_\ell \sum_{\tilde{\boldsymbol{y}} \in \mathcal{S}_l} \int \Pr(\boldsymbol{Y}_i = \tilde{\boldsymbol{y}} | \boldsymbol{b}, \boldsymbol{X}_i, \boldsymbol{W}_i, \boldsymbol{U}_i) dG(\boldsymbol{b}; \boldsymbol{\theta})} \right\} \ . \qquad (3.31)$$

Under the GLMM assumptions for correlated binary data, (3.31) can be written as

$$L(\boldsymbol{\alpha},\boldsymbol{\theta}) = \prod_{i=1}^{K} \left\{ \frac{\pi(\boldsymbol{y}_i) \int \prod_{j=1}^{n_i} \left[\mu_{ij}^c\right]^{y_{ij}} \left(1-\mu_{ij}^c\right)^{1-y_{ij}} dG(\boldsymbol{b};\boldsymbol{\theta})}{\sum_{\ell=1}^{R} \pi_\ell \sum_{\tilde{\boldsymbol{y}}\in\mathcal{S}_\ell} \int \prod_{j=1}^{n_i} \left[\mu_{ij}^c\right]^{\tilde{y}_{ij}} (1-\mu_{ij}^c)^{1-\tilde{y}_{ij}} dG(\boldsymbol{b};\boldsymbol{\theta})} \right\} \tag{3.32}$$

In practice, maximizing (3.32) is no trivial task – even if the sampling rates are known. It involves the following steps:

- enumeration of the possible values of $\boldsymbol{Y}_i$ in the denominator
- marginalization through integration in both the numerator and denominator
- maximization of the marginal likelihood

When the sampling rates are not known, direct maximization of (3.32) is not possible. In this situation, the unknown sampling intensities are said to be nuisance parameters and they must be dealt with in some manner.

Neuhaus and McCulloch (2006) suggested that, for GLMMs, the biased sampling problem is connected to misspecification of the mixing distribution in the following way. Re-write (3.30) as

$$\Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i | S_i = 1, \boldsymbol{Z}_i) = \int \Pr(\boldsymbol{Y}_i = \boldsymbol{y}_i | \boldsymbol{b}, \boldsymbol{Z}_i, \boldsymbol{U}_i) dG^*(\boldsymbol{b}, \boldsymbol{Z}_i, \boldsymbol{y}_i; \boldsymbol{\pi}, \boldsymbol{\theta}) \tag{3.33}$$

where

$$dG^*(\boldsymbol{b}, \boldsymbol{Z}_i, \boldsymbol{y}_i; \boldsymbol{\pi}, \boldsymbol{\theta}) := \frac{\pi(\boldsymbol{y}_i) dG(\boldsymbol{b};\boldsymbol{\theta})}{\sum_{\ell=1}^{R} \pi_\ell \sum_{\tilde{\boldsymbol{y}}\in\mathcal{S}_\ell} \int \Pr(\boldsymbol{Y}_i = \tilde{\boldsymbol{y}} | \boldsymbol{b}, \boldsymbol{Z}_i, \boldsymbol{U}_i) dG(\boldsymbol{b};\boldsymbol{\theta})} \ . \tag{3.34}$$

By interpreting $dG^*(\boldsymbol{b}, \boldsymbol{Z}_i, \boldsymbol{y}_i; \boldsymbol{\pi}, \boldsymbol{\theta})$ as a new random-effects density, an analysis that ignores the sampling scheme misspecifies the mixing distribution as G instead of G^*. In other words, the sample model induced by the sampling scheme can be written as

$$h(\nu_{ij}^c) = \alpha_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_1 + \boldsymbol{w}_i^\top \boldsymbol{\alpha}_2 + \boldsymbol{u}_{ij}^\top \boldsymbol{b}_i^* \tag{3.35}$$

where $\nu_{ij}^c = E(Y_{ij} | S_i = 1, \boldsymbol{x}_{ij}, \boldsymbol{w}_i, \boldsymbol{u}_{ij})$, and $\boldsymbol{b}_i^* \sim G^*$. A consequence of (3.34) is that, unlike $\boldsymbol{b}_i$ in (3.28), $\boldsymbol{b}_i^*$ in (3.35) is correlated with $\boldsymbol{Z}_i$ which is a violation of one of the GLMM assumptions. Neuhaus and McCulloch (2006) showed that when the random effects are correlated with the

covariates, maximum likelihood estimates of $\boldsymbol{\alpha}$ are biased if the correlation is ignored.

In the sections to follow, we consider a random intercepts model and show that $\boldsymbol{\alpha}_1$ can be consistently estimated by (2.28) when sampling of clusters is performed based on the total.

3.10.1 Sampling Clusters Based on the Number of Events

Assume the following random intercepts GLMM:

$$h(\mu_{ij}^c) = b_i + \alpha_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_1 + \boldsymbol{w}_i^\top \boldsymbol{\alpha}_2 \tag{3.36}$$

where $b_i \sim N(0, \sigma_b^2)$. Assume that clusters are sampled based on (3.12) which we reproduce here for convenience:

$$\pi(\boldsymbol{y}_i) = \begin{cases} \pi_1 & \text{if } t_i \leq m \\ \pi_2 & \text{if } t_i > m \end{cases}$$

From (3.35), the sample model induced by this sampling scheme is

$$h(\nu_{ij}^c) = b_i^* + \alpha_0 + \boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_1 + \boldsymbol{w}_i^\top \boldsymbol{\alpha}_2 \tag{3.37}$$

where the new random effects density is

$$dG^*(b, \boldsymbol{Z}_i, y_i; \boldsymbol{\pi}_i, \sigma_b^2) = \frac{\pi(\boldsymbol{y}_i) dG(b; \sigma_b^2)}{\pi_1 \sum_{\tilde{\boldsymbol{y}} \in \mathcal{S}_1} \Pr(Y_i = \tilde{\boldsymbol{y}} | \boldsymbol{Z}_i) + \pi_2 \sum_{\tilde{\boldsymbol{y}} \in \mathcal{S}_2} \Pr(Y_i = \tilde{\boldsymbol{y}} | \boldsymbol{Z}_i)}$$

and

- $b_i^* \sim G^*$,
- $\mathcal{S}_1 = \{(y_1, y_2, \ldots, y_{n_i}) : t_i \leq m\}$,
- $\mathcal{S}_2 = \{(y_1, y_2, \ldots, y_{n_i}) : t_i > m\}$,

Assume h is the log link and write (3.37) as

$$\log(\nu_{ij}^c) = \delta_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\alpha}_1 \tag{3.38}$$

where $\delta_i = b_i^* + \alpha_0 + \boldsymbol{w}_i^\top \boldsymbol{\alpha}_2$. Since (3.38) has the same structure as (2.15), $\boldsymbol{\alpha}_1$ can be estimated as the root of the estimating function given in (2.28). Robust standard errors of $\widehat{\boldsymbol{\alpha}}_1$ can be approximated by (2.31).

3.10.2 Sampling Clusters Exhibiting Variation

As in the previous section, assume the same random intercepts GLMM. Suppose the sampling scheme samples clusters based on whether the response vector exhibits variation. Recall from section 3.7.2 that this sampling scheme can be described by (3.19) which is reproduced here for convenience:

$$\pi(\boldsymbol{y}_i) = \begin{cases} \pi_1 & \text{if } t_i = 0 \text{ or } t_i = n_i \\ \pi_2 & \text{if } 0 < t_i < n_i \end{cases}$$

The sample model is exactly the same as (3.37) except that under this sampling scheme,

$$dG^*(b, \boldsymbol{Z}_i, \boldsymbol{y}_i; \boldsymbol{\pi}, \sigma_b^2) = \frac{\pi(\boldsymbol{y}_i) dG(b; \sigma_b^2)}{(r_{12} - 1)\{\Pr(\boldsymbol{Y}_i = \boldsymbol{0}|\boldsymbol{Z}_i) + \Pr(\boldsymbol{Y}_i = \boldsymbol{1}|\boldsymbol{Z}_i)\} + 1}$$

where $r_{12} = \pi_1/\pi_2$. If we assume the log link, then the sample model is the same as (3.38) in which $\boldsymbol{\alpha}_1$ can be consistently estimated by (2.28).

3.11 Simulations: GLMMs and Outcome-Dependent Sampling

In this section, the result of a simulation study to assess the performance of our method in the context of GLMMs is presented. We consider correlated binary outcomes with cluster size 5 and assume that the relationship between the outcomes and covariates in the population is described by

$$h(\mu_{ij}^c) = b_i + \alpha_0 + x_{ij}\alpha_1$$

where $b_i \sim N(0, \sigma^2)$, $\mu_{ij}^c = \Pr(Y_{ij} = 1|b_i, x_{ij})$ with $X_{ij} \sim \text{Bernoulli}(p)$. Two link functions are considered: log and logit. We assume the following outcome-dependent sampling scheme:

$$\pi(\boldsymbol{y}_i) = \begin{cases} 0 & \text{if } t_i = 0, \\ 1 & \text{if } t_i > 0 \end{cases}$$

Selection stops when there are K clusters in the sample with at least one diseased member. This is analogous to sampling N_K clusters until we obtain K clusters with at least 1 diseased member and using only these K clusters in the analysis.

3.11.1 Log Link

For the log link, we assume that the population model is described by

$$\log \mu_{ij}^c = b_i + \alpha_0 + \alpha_1 x_{ij}$$

It is assumed that $\sigma_b^2 = 1$, $\alpha_0 = -5.0$ and $\alpha_1 \in \{0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0\}$. Under the assumption that $X \sim \text{Bernoulli}(p)$, these chosen parameter values correspond roughly to prevalence of up to 0.05. This is obtained by considering

$$\begin{aligned} \Pr(Y = 1) &= E_X[\Pr(Y = 1|X)] = E_X\Big[E_b[\Pr(Y = 1|b, X)]\Big] \\ &= E_X[e^{\alpha_0 + \alpha_1 X + \sigma_b^2/2}] = e^{-\alpha_0 + \sigma_b^2/2} E_X[e^{\alpha_1 X}] \\ &= e^{-\alpha_0 + \sigma_b^2/2} \left(\frac{1 + e^{\alpha_1}}{2}\right) \end{aligned}$$

We sample until $K = 100$ clusters with $t_i > 0$ is obtained.

Table 3.10 shows that on average, our estimating function provides consistent estimates of the slope α_1. The robust standard errors are different from the true standard errors but, overall, they are relatively close to each other. Coverage is close to the nominal value of 95%. This suggests that our estimating function performs well in this biased sample setup.

Table 3.10: 2000 simulations were performed. N_K is the average number of clusters that we needed to sample in order to obtain 100 clusters that have at least 1 diseased member. Estimation is done using the estimating function (2.28)

α_1	p	N_K	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Robust SE($\hat{\alpha}_1$)	Coverage
2.0	0.5	536	2.02	0.3081	0.2969	0.950
1.8	0.5	616	1.82	0.2895	0.2797	0.943
1.6	0.5	710	1.62	0.2745	0.2653	0.940
1.4	0.5	818	1.41	0.2522	0.2518	0.952
1.2	0.5	939	1.22	0.2464	0.2414	0.942
1.0	0.5	1071	1.01	0.2309	0.2321	0.952
0.8	0.5	1230	0.80	0.2234	0.2250	0.948
0.6	0.5	1383	0.61	0.2224	0.2195	0.951
0.4	0.5	1554	0.40	0.2191	0.2157	0.941

Table 3.11: 2000 simulations were performed. N_K is the average number of clusters that we needed to sample in order to obtain 100 clusters that have at least 1 diseased member. Estimation is done using between-within decomposition

α_1	p	N_K	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Likelihood-Based SE($\hat{\alpha}_1$)	Coverage
2.0	0.5	536	1.99	0.3342	0.2747	0.926
1.8	0.5	616	1.79	0.2818	0.2572	0.929
1.6	0.5	710	1.60	0.2661	0.2420	0.932
1.4	0.5	818	1.39	0.2504	0.2305	0.934
1.2	0.5	939	1.20	0.2401	0.2183	0.925
1.0	0.5	1071	0.99	0.2250	0.2094	0.931
0.8	0.5	1230	0.79	0.2196	0.2029	0.929
0.6	0.5	1383	0.60	0.2199	0.1981	0.924
0.4	0.5	1554	0.40	0.2174	0.1949	0.923

Table 3.11 suggests that the method of between-within covariate decomposition (Neuhaus and McCulloch, 2006) also provides reasonable estimates of the slope parameter. As with our estimating function, estimates of the slope based on this method are, on average, very close to the true values. Note, however, that the likelihood-based standard errors are consistently smaller than the actual standard errors. The coverage is also smaller than the nominal 95% value.

Table 3.12 provides estimates when a GLMM is fit to the biased sample ignoring the sam-

pling scheme. For all values of α_1, note that $\hat{\alpha}_1$ are all biased. An interesting feature of this table is that when the true slope is 0.4, the likelihood-based standard error is noticeably larger than the true standard error (almost twice as large).

Table 3.12: 2000 simulations were performed. N_K is the average number of clusters that we needed to sample in order to obtain 100 clusters that have at least 1 diseased member. Estimation is done by naively fitting a mixed model to the biased data

α_1	p	N_K	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Likelihood-Based SE($\hat{\alpha}_1$)	Coverage
2.0	0.5	536	1.77	0.3009	0.2701	0.807
1.8	0.5	616	1.58	0.2613	0.2514	0.815
1.6	0.5	710	1.39	0.2409	0.2347	0.814
1.4	0.5	818	1.20	0.2190	0.2201	0.812
1.2	0.5	939	1.02	0.2084	0.2074	0.831
1.0	0.5	1071	0.84	0.1936	0.1963	0.853
0.8	0.5	1230	0.67	0.1852	0.1877	0.881
0.6	0.5	1383	0.50	0.1835	0.1812	0.900
0.4	0.5	1554	0.33	0.1793	0.3308	0.925

3.11.2 Logit Link

In this subsection, we assume that the population model is described by

$$\text{logit}\,\mu_{ij}^c = \eta_{ij}$$

We consider three models:

$$\text{Model 1: } \eta_{ij} = b_i + \alpha_0 + x_{ij}\alpha_1$$

$$\text{Model 2: } \eta_{ij} = b_i + \alpha_0 + x_{ij}\alpha_1 + c_i\alpha_2$$

$$\text{Model 3: } \eta_{ij} = b_i + \alpha_0 + x_{ij}\alpha_1 + c_i\alpha_2 + c_i x_{ij}\alpha_3$$

For all three models, we assume that $b_i \sim N(0, \sigma_b^2)$.

Model 1

As with the previous section, the sampling scheme samples until we obtain $K = 100$ clusters with at least 1 diseased member. Data analysis is performed only on these 100 clusters. It is also assumed that:

- $\sigma_b^2 = 1$
- $n_i = 5$ for all i
- $\alpha_0 \in \{-3, -4, -5, -6\}$, $\alpha_1 \in \{0.5, 1.0\}$
- $X_{ij} \sim \text{Bernoulli}(p = 0.5)$

Table 3.13: 2000 simulations were performed. Estimation technique is performed using our estimating function

α_0	α_1	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Robust SE($\hat{\alpha}_1$)	Coverage
-6	0.5	0.50	0.2271	0.2259	0.946
-5	0.5	0.49	0.2183	0.2185	0.949
-4	0.5	0.47	0.2094	0.2043	0.945
-3	0.5	0.42	0.1836	0.1804	0.920
-6	1.0	1.00	0.2455	0.2432	0.952
-5	1.0	0.96	0.2335	0.2322	0.941
-4	1.0	0.91	0.2148	0.2133	0.915
-3	1.0	0.82	0.1818	0.1835	0.810

Table 3.13 shows that only under rare disease can we use our estimating function to obtain consistent estimates of the slope. Here, the phrase "rare disease" is defined by the following combinations of (α_0, α_1): $(-6, 0.5)$, $(-6, 1)$, $(-5, 0.5)$, and $(-5, 1)$. For these values, the robust standard errors are approximately equal to the true standard errors and coverage is approximately equal to the nominal value of 95%.

To these simulated values we also fit a mixed model using the between-within decomposition procedure. The results are provided in table 3.14. Unlike the log link, the procedure is not effective for the logit link. For all combinations of α_0 and α_1, estimates of α_1 are biased.

Table 3.14: 2000 simulations were performed. Estimation technique is performed using between-within decomposition

α_0	α_1	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Likelihood-Based SE($\hat{\alpha}_1$)	Coverage
-6	0.50	0.63	0.2844	0.2549	0.890
-5	0.50	0.62	0.2764	0.2508	0.907
-4	0.50	0.60	0.2717	0.2432	0.905
-3	0.50	0.58	0.2501	0.2304	0.918
-6	1.00	1.25	0.3031	0.2731	0.839
-5	1.00	1.21	0.2926	0.2659	0.869
-4	1.00	1.19	0.2764	0.2548	0.879
-3	1.00	1.13	0.2466	0.2375	0.916

For comparison, we also fit Model 1 to the biased data without adjusting for the sampling scheme. Results are presented in table 3.15. The results are not what we expected. Unlike the case where the true model is based on the log link, the case where the true model is based on the logit link seems to contradict our earlier discussion about biased sampling. All estimates of α_1 are not biased. The performance is better than our estimating function technique. To try to understand what is happening, we also simulated data under the assumption that $\sigma_b^2 = 0.3,\ 4.0,\ 9.0$. The results are not shown but a quick summary is as follows.

- When $\sigma_b^2 = 0.3$, the naive approach continues to perform well
- When $\sigma_b^2 = 4.0$, the naive estimates are slightly biased
- When $\sigma_b^2 = 9.0$, the naive estimates are biased

This seems to suggest that if the population of clusters is very heterogeneous (as defined by a large variance component), then naive fitting of the biased sample leads to biased estimates. This is a conjecture; we do not have any proof that this is indeed the case.

In an early article, Neuhaus and Jewell (1990) discussed biased sampling under the logit-normal model and suggested that if the sampling scheme is not taken into account, parameter estimates are biased. The simulation results seem to contradict their conclusions. They provided an example where the regression model contains a single cluster-constant covariate and showed analytically that the estimate of the parameter associated with that covariate is biased. As

suggested by the theory, our simulation also shows that all parameters associated with cluster-constant covariates are biased. However, parameters associated with cluster-varying covariates seem consistent. Similar behavior are also seen under models 2 and 3 (see below). At this point, we do not have an answer to why this is the case. There are several questions that need to be answered (to be addressed in the future). First, does this behavior continue to hold under multivariate random effects structures? Second, if the true mixing distribution is correlated with covariates, then under our biased sampling setup, will naive estimation continue to perform better than our estimating function under the rare disease scenario?

Table 3.15: 2000 simulations were performed. Estimates are obtained by naively fitting a GLMM without adjusting for sampling scheme

α_0	α_1	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Likelihood-Based SE($\hat{\alpha}_1$)	Coverage
-6	0.5	0.51	0.2300	0.2280	0.946
-5	0.5	0.50	0.2237	0.2241	0.960
-4	0.5	0.50	0.2207	0.2172	0.951
-3	0.5	0.49	0.2095	0.2058	0.947
-6	1.0	1.02	0.2465	0.2445	0.949
-5	1.0	1.00	0.2401	0.2379	0.950
-4	1.0	0.99	0.2290	0.2274	0.947
-3	1.0	0.97	0.2096	0.2115	0.955

Model 2

For model 2, we generated clustered data based on the assumptions that $n_i = 5$, $\sigma_b^2 = 1$, $\alpha_0 = -6$, $\alpha_1 \in \{0, 0.4, 0.6, 0.8, \ldots, 2.0\}$, $\alpha_2 = 1$, $X_{ij} \sim$ Bernoulli(0.5), and $C_i \sim$ Bernoulli(0.5). We sample until we obtained $K = 100$ clusters with at least one diseased member.

Table 3.16 provides estimates using our estimating function. For α_1 values of 1.4 or less, our estimating function provides reasonably good estimates. Coverage is approximately equal to the nominal value of 95%.

Table 3.17 provides estimates using the between-within decomposition procedure. As with Model 1, this method is not effective.

Table 3.16: 2000 simulations were performed based on Model 2. N_K is the average number of clusters that we needed to sample in order to obtain 100 clusters that have at least 1 diseased member. Biased data is estimated using our estimating function

α_1	N_K	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Likelihood SE($\hat{\alpha}_1$)	Coverage
2.0	813	1.90	0.3134	0.2974	0.912
1.8	934	1.73	0.2925	0.2829	0.930
1.6	1070	1.54	0.2810	0.2673	0.925
1.4	1224	1.36	0.2606	0.2552	0.946
1.2	1340	1.17	0.2428	0.2443	0.945
1.0	1602	0.98	0.2321	0.2359	0.955
0.8	1815	0.78	0.2237	0.2279	0.947
0.6	2051	0.59	0.2269	0.2226	0.942
0.4	2299	0.40	0.2188	0.2191	0.957
0.0	2821	0.00	0.2219	0.2164	0.948

Table 3.17: 2000 simulations were performed based on Model 2. Biased data is estimated using covariate centering

α_1	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Likelihood SE($\hat{\alpha}_1$)	Coverage
2.0	2.37	0.3672	0.3300	0.817
1.8	2.17	0.3451	0.3157	0.801
1.6	1.93	0.3391	0.3001	0.811
1.4	1.71	0.3198	0.2886	0.818
1.2	1.47	0.2992	0.2775	0.836
1.0	1.24	0.2892	0.2684	0.847
0.8	0.99	0.2809	0.2604	0.877
0.6	0.75	0.2857	0.2545	0.893
0.4	0.50	0.2764	0.2501	0.911
0.0	-0.01	0.2805	0.2471	0.918

Table 3.18 provides results in which a GLMM is fit to the data without adjusting for the sampling scheme. As with Model 1, the naive fit performs better than our procedure. Note that the naive method provides biased estimates of α_2 - the parameter associated with the cluster-constant covariate c.

Table 3.18: 2000 simulations were performed based on Model 2. Estimation is performed by naively fitting a GLMM on the biased data

α_1	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Likelihood SE($\hat{\alpha}_1$)	Coverage
2.0	2.00	0.3136	0.3004	0.950
1.8	1.82	0.2941	0.2862	0.950
1.6	1.61	0.2832	0.2713	0.944
1.4	1.41	0.2643	0.2594	0.952
1.2	1.21	0.2480	0.2489	0.953
1.0	1.02	0.2365	0.2404	0.958
0.8	0.81	0.2291	0.2329	0.955
0.6	0.61	0.2319	0.2278	0.948
0.4	0.41	0.2220	0.2239	0.956
0.0	0.00	0.2256	0.2212	0.949

Model 3

For this model, we assume the following: $n_i = 5$, $\sigma_b^2 = 1$, $\alpha_0 = -6$, $\alpha_2 = 1$, $\alpha_3 = 0.75$, $\alpha_1 \in \{0, 0.4, 0.6, \ldots, 2.0\}$, $X_{ij} \sim$ Bernoulli(0.5), and $C_i \sim$ Bernoulli(0.5).

Table 3.19 provides estimates using our estimating function. While our method provides good estimates of α_1, the parameter of interest is α_3 when the model contains an interaction term. It is clear from the simulation results that our method produces biased estimates of α_3 except when $\alpha_3 = 0$. Increasing the sample size from 100 to 300 does not help.

Table 3.20 provides estimates from naively fitting Model 3 to the biased sample without adjusting for the sampling scheme. The naive method seems to produce better estimates of α_3 then our method, at least for values of α_1 up to 0.8. However, estimates of the interaction term are also poor except where the magnitude of α_1 is small.

3.12 Summary of Simulation Results

The simulation study provided under the marginal model and GLMM setup leads to both expected and unexpected results.

First, under marginal regression models for correlated binary data, if the assumed relationship between the outcome and covariates is described by the log link, then estimates of the

slope based on our estimating function are consistent for the biased sampling scheme that we considered. For rare outcomes, our procedure is still useful for providing consistent estimates of the odds ratio even when the assumed relationship between the outcome and covariates is described by the logit link. Confidence intervals constructed using the sandwich estimator leads to coverage that is near the nominal value. Naive fitting using GEE leads to biased results for both the log and logit link functions. The simulation study also shows that there are substantial efficiency gains, especially under the rare disease assumption considered by logistic regression.

For GLMMs with a random intercept, if the assumed relationship between the outcome and covariates is described by the log link, then our estimating function provides consistent estimates of the parameters associated with cluster-varying covariates. The between-within covariate decomposition method of Neuhaus and McCulloch (2006) seems to provide consistent estimates of the slope but has poorer coverage. Naively fitting a GLMM under the biased sampling scheme considered here is not competitive since estimates are not consistent.

When the specified model is based on the logit link and the outcome is rare, our procedure continues to provide consistent estimates for the slope when the true model does not contain interaction terms. However, naive fitting provides better estimates then our procedure with respect to consistency. When the true model contains interaction terms, our procedure breaks down entirely in the sense that estimates of the interaction parameter is very biased. The between-within approach also performs very poorly. Naively fitting a GLMM ignoring the sampling scheme does not help either. In this scenario, the only feasible approach may be conditional logistic regression as it is unaffected by our sampling scheme (Neuhaus and Jewell, 1990).

3.13 Discussion

The estimation procedure that we have adopted for variable probability sampling based on the total seems to work reasonably well when the outcome is related to the covariates through the log link function. What is remarkable is that the procedure works whether the specified model is a marginal model or a mixed model. This is perhaps not surprising in light of the fact

that, for a random intercept GLMMs with log link function, the slope parameter is the same as an equivalent marginal model. Furthermore, our procedure provides consistent estimates of parameters associated with cluster-varying covariates while making no distinction between a model that contains cluster-constant covariates and one that does not. In this sense, we may interpret our procedure as always adjusting for cluster-constant covariates.

Variable probability sampling is relevant when it is relatively cheap or easy to obtain outcome information about each cluster and when it is expensive to measure covariates. One could imagine a study where obtaining the disease status is operationally cheap but obtaining genetic information is expensive. A case-control study examining the association between a disease and genetic risk factors can benefit by selecting sibling controls. By viewing the case and his/her siblings as a cluster, such a design picks only informative clusters (informative with respect to our estimation procedure). This is equivalent to our biased sampling framework where sibships with no diseased member are not sampled (truncated sampling). The advantages and disadvantages of using sibling controls are thoroughly-discussed by Wacholder et al. (1992a,b). Our estimation procedure was derived under the assumption that the outcome and covariates are described by a generalized linear model with log link. While this may be a limitation, it is also its strength because it provides a procedure for estimating risk ratios under case-control sampling without resorting to auxiliary information that may or may not be available. In epidemiologic studies, risk-ratio modeling is usually not done for case-control designs because without auxiliary information, the risk ratio cannot be estimated. Hence, the popularity of odds-ratio modeling and logistic regression. Our estimation procedure recognizes the correlation between siblings whereas the usual case-control analysis under logistic regression does not. However, under the rare disease assumption, we can use our procedure to estimate the slope for a logistic regression model. We have shown in our simulation study (see Table 3.9) that when there is correlation among siblings, logistic regression under truncated sampling is biased but our procedure approximates marginal odds ratios well (assuming the disease is rare). This raises the question of validity for case-control studies using siblings as controls whereby the method of analysis is prospective maximum likelihood and the model is a logistic regression model. For matched studies with correlated binary outcomes, Liang (1987) suggested the use

of a generalized Mantel-Haenszel procedure.

We have yet to understand why naive fitting under the logit-normal GLMM leads to consistent estimates of the slope when the true model contains no interaction terms. There seems to be a connection to the results of Prentice and Pyke (1979) but we have no formal way of showing that this is the case or whether this is even true in general. Our simulation suggests that when the true model contains interactions between cluster-constant and cluster-varying covariates, even naive fitting is unable to consistently estimate the interaction parameters. A natural question to ask is whether lack of consistency is also true for interactions involving two cluster-varying covariates or whether the problem goes away under large samples. These are very interesting questions whose answers we hope to provide in forthcoming research.

Our procedure assumes that the biased sample was obtained via variable probability sampling. Similar biased samples can be obtained by the standard stratified sampling scheme as discussed in Section 3.4.2. In a fully parametric setup, Neuhaus et al. (2002, 2006) considered the situation where the biased sample was obtained by stratified sampling but used procedures developed under variable probability sampling to obtain estimates. Based on their results, we conjecture that our estimating function is also applicable even when the sample is obtained by stratified sampling, at least for the purpose of obtaining consistent estimates. This is something to be examined in the future.

Implicit in our procedure is the assumption of no interference. In the language of Diggle et al. (2002, pg. 255), this is also known as the *full covariate conditional mean* assumption. For the double-pair design considered earlier, risk of death for an occupant may be related to the seat belt status of another occupant. For example, during an accident, an occupant who is not secured by a seat belt may be tossed in such a way as to cause harm or death to another occupant. In cases where there is interference, it is unclear whether our estimating function continues to provide consistent estimates. For GEE, Pepe and Anderson (1994) suggested that working independence should be used when there is interference. Our estimating function was a modification of GEE under working independence. By analogy, we conjecture that it should perform reasonably well, even when there is interference.

Table 3.19: 2000 simulations were performed based on Model 3. Biased sample is fitted using our estimating function.

α_1	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Likelihood SE($\hat{\alpha}_1$)	Coverage	$\hat{\alpha}_3$	True SE($\hat{\alpha}_3$)	Likelihood SE($\hat{\alpha}_3$)	Coverage
				$K = 100$				
2.0	1.94	0.6356	0.7083	0.948	0.60	0.7597	0.8263	0.9780
1.8	1.80	0.6495	0.6825	0.955	0.59	0.7641	0.7933	0.9750
1.6	1.62	0.6459	0.6534	0.954	0.57	0.7299	0.7525	0.9730
1.4	1.44	0.6459	0.6306	0.955	0.59	0.7439	0.7226	0.9620
1.2	1.24	0.6106	0.5939	0.952	0.62	0.7003	0.6824	0.9605
1.0	1.07	0.5968	0.5766	0.959	0.61	0.6859	0.6569	0.9610
0.8	0.84	0.5829	0.5499	0.948	0.66	0.6624	0.6247	0.9460
0.6	0.62	0.5533	0.5344	0.965	0.69	0.6248	0.6053	0.9550
0.4	0.43	0.5443	0.5181	0.952	0.69	0.6066	0.5856	0.9550
0.0	0.01	0.5199	0.4955	0.951	0.71	0.5702	0.5572	0.9540
				$K = 300$				
2.0	1.99	0.4265	0.4046	0.939	0.51	0.4816	0.4668	0.955
1.8	1.80	0.4067	0.3884	0.943	0.53	0.4622	0.4414	0.944
1.6	1.59	0.3620	0.3637	0.949	0.58	0.4192	0.4181	0.947
1.4	1.41	0.3697	0.3479	0.949	0.59	0.4148	0.3986	0.944
1.2	1.21	0.3271	0.3315	0.962	0.62	0.3746	0.3795	0.949
1.0	1.01	0.3381	0.3186	0.941	0.64	0.3827	0.3637	0.943
0.8	0.80	0.3112	0.3077	0.952	0.67	0.3551	0.3503	0.949
0.6	0.60	0.2978	0.2995	0.953	0.68	0.3402	0.3398	0.950
0.4	0.42	0.2929	0.2916	0.955	0.68	0.3306	0.3298	0.950
0.0	-0.01	0.2781	0.2831	0.960	0.73	0.3126	0.3185	0.959

Table 3.20: 2000 simulations were performed based on Model 3. Biased sample is naively fitted

α_1	$\hat{\alpha}_1$	True SE($\hat{\alpha}_1$)	Likelihood SE($\hat{\alpha}_1$)	Coverage	$\hat{\alpha}_3$	True SE($\hat{\alpha}_3$)	Likelihood SE($\hat{\alpha}_3$)	Coverage
				$K = 100$				
2.0	2.86	3.5699	23.7335	0.937	-0.11	3.5860	23.8453	0.948
1.8	2.44	3.0300	11.5624	0.947	0.12	3.0687	11.6680	0.946
1.6	2.03	2.4332	10.1609	0.955	0.33	2.4859	10.2577	0.958
1.4	1.72	2.0963	7.5628	0.953	0.44	2.1251	7.6537	0.958
1.2	1.43	1.6916	5.5668	0.961	0.52	1.7243	5.6522	0.961
1.0	1.13	1.0164	1.0702	0.968	0.63	1.0711	1.1505	0.957
0.8	0.87	0.8399	1.8701	0.957	0.69	0.9018	1.9464	0.952
0.6	0.64	0.6650	0.9897	0.967	0.73	0.7219	1.0617	0.961
0.4	0.45	0.6705	0.5276	0.960	0.72	0.7237	0.5961	0.956
0.0	0.01	0.5208	0.5068	0.960	0.73	0.5740	0.5699	0.963
				$K = 300$				
2.0	2.05	0.5411	0.4021	0.952	0.66	0.5822	0.4653	0.960
1.8	1.84	0.4030	0.3826	0.955	0.67	0.4594	0.4415	0.948
1.6	1.63	0.3642	0.3639	0.957	0.69	0.4207	0.4192	0.960
1.4	1.44	0.3667	0.3490	0.952	0.69	0.4145	0.4008	0.955
1.2	1.23	0.3261	0.3329	0.961	0.71	0.3771	0.3819	0.957
1.0	1.03	0.3403	0.3215	0.945	0.72	0.3861	0.3676	0.946
0.8	0.81	0.3118	0.3103	0.956	0.74	0.3584	0.3539	0.953
0.6	0.61	0.2993	0.3022	0.958	0.73	0.3439	0.3435	0.951
0.4	0.43	0.2942	0.2951	0.952	0.72	0.3330	0.3344	0.952
0.0	-0.01	0.2816	0.2865	0.963	0.75	0.3163	0.3229	0.962

Chapter 4

Applications

4.1 Introduction

In chapter 1, we studied estimating functions in the presence of cluster-specific nuisance parameters. When marginal models contain cluster-specific intercepts and when the mixing distribution is misspecified for random intercept GLMMs, standard approaches to estimation may fail to provide consistent estimates while our estimating function produces consistent estimates of parameters associated with cluster-varying covariates. Chapter 2 considers biased sampling in the context of variable probability sampling based on cluster totals and casts the biased sampling problem as a nuisance parameter problem. This enabled us to use results established in chapter 1 to provide consistent estimates of the slope when the link is the log function.

In this chapter, we apply these established results to real data. Section 4.2 applies our estimating function to two biased samples of twin pairs from a large data set. Efficiency gains through the use of biased sampling schemes together with our estimation procedure is illustrated. Section 4.3 implements our procedure to a truncated data set used in the study of helmet effectiveness. A comparison is made between our method and an existing procedure for matched pairs data. Section 4.4 applies our procedure to Phase I and II of the Carolina Breast Cancer Study.

4.2 Birth Weight and Risk of Death

Ananth et al. (2003) examined a relationship between perinatal outcomes and birth weight among twin births in the United States. They used the 1995 to 1997 matched multiple birth file compiled by the Division of Vital Statistics in the National Center for Health Statistics. This example uses the 1995 to 2000 multiple birth files which contain 286358 pairs of twins (see Table A.1). We assumed that the 286358 pairs is a random sample of twin pairs from an infinite population in which the relationship between death and birth weight is described by the marginal model

$$\log \mu_{ij} = \alpha_0 + \alpha_1 x_{ij} \ , \tag{4.1}$$

where $\mu_{ij} = E(Y_{ij}|\boldsymbol{X}_i) = \Pr(Y_{ij} = 1|\boldsymbol{x}_{ij})$, Y_{ij} is a 0/1 binary random variable whose value is 1 if the j-th infant from the i-th twin pair died within the first year of birth, and $\boldsymbol{x}_{ij}$ is birth weight (in lbs.). Throughout, we refer to this data as the "full data".

Using the full data, parameters were estimated using GEE assuming independence working correlation and Bernoulli variance function $v_{ij} = \mu_{ij}(1-\mu_{ij})$. Independence working correlation was used because our estimating function approach uses a sub-optimal weight based on the independence working correlation. This facilitates comparisons with estimates obtained from our approach. The estimated slope and its associated standard error are $-0.5611(0.0203)$.

We illustrate the use of our estimating function and the benefit of biased sampling by considering first the following biased sampling scheme:

$$\pi(\boldsymbol{y}_i) = \begin{cases} \pi_1 = 0.01 & \text{if } t_i = 0 \\ \pi_2 = 1 & \text{if } t_i > 0 \end{cases} \tag{4.2}$$

where t_i is the total number of deaths in the i-th pair. In particular, this scheme samples from the 286358 twin pairs all pairs with at least one death and only one percent of the other pairs. As shown in chapter 3, this sampling scheme induces the sample model

$$\log \nu_{ij} = \delta_i + \alpha_1 x_{ij} \tag{4.3}$$

where $\nu_{ij} = E[Y_{ij} \mid S_i = 1, x_{ij}]$, and $\delta_i = \alpha_0 - \log\{1 + \Pr(T_i = 0; \boldsymbol{X}_i)(\pi_1/\pi_2 - 1)\}$. Table A.2 provides a brief description of a sample after applying (4.2). We estimate α_1 using both naive GEE (without taking into account the sampling scheme) and our estimating function. Table A.3 compares estimates of the slope between naive GEE and our estimating function. While the standard error based on naive GEE is small, the estimate of the slope is biased. On the other hand, the estimate obtained from our estimating function is much closer to the value based on the full data: -0.4798 versus -0.5611. All things being equal, we expect the efficiency ratio between the estimate using all pairs and the estimate using 6036 pairs to be approximately 47.44 (286358/6036). However, the actual ratio is 3.54 ($(0.0382/0.0203)^2$). This implies that this biased sampling design is more efficient than simple random sampling – the efficiency increases by a factor of 3.5 relative to a simple random sample of the same size. It also implies that a random sample of size greater than 81000 pairs is needed to achieve the same amount of efficiency as our biased sample of only 6036 pairs.

It may be argued that our estimate of -0.4798 based on a biased sample of 6036 pairs is not close to the estimate of -0.5611 based on the full data set. Recall however that we assumed the true model contains only a single covariate, birth weight. But the true model is unknown and may contain other covariates. As discussed in the concluding remarks of chapter 3, our procedure always adjusts for cluster-level covariates. For the analysis based on the full data, we fit a model that contains only birth weight in the linear predictor. The true model may contain other cluster-level covariates that we did not specify. For example, these potentially informative cluster-level covariates might include mothers' smoking activity, quantity of alcohol consumed, or number of prenatal care visits to the doctor's office during pregnancy. If the true model contains additional cluster-level covariates, our estimating function returns an unbiased estimate of α_1 adjusting for these other covariates. Motivated by this, we extended model (4.1) to include the following additional cluster-level covariates: gestational age (binary: 0 if less than 36 weeks and 1 if at least 36 weeks), number of prenatal care visits to the doctor's

office (binary: 0 if less than 12 times and 1 if at least 12 times), gender, and mother's marital status. Using the full data set, we fit this model using GEE (not shown) with independence working correlation and Bernoulli variance function. The estimated risk ratio (in the log scale) associated with birth weight shrinks to $-0.5273(0.0260)$, a value closer to -0.4798.

Note that the sample model (4.3) has the structure of a GLMM if δ_i is viewed as a random intercept. For the purpose of comparison, we fit a GLMM to (4.3) using the method of between-within covariate decomposition (Neuhaus and McCulloch, 2006). The rationale for choosing this method is because δ_i is a function of the covariates (refer to section 3.7.1). We assumed that $\delta_i \sim N(0, \sigma_\delta^2)$. The NLMIXED procedure (SAS Institute Inc, 2009) was used to estimate α_1. The column labeled "Between-Within" in Table A.3 shows that this method, at least for this sample, provides a reasonable estimate of the effect associated with birth weight: $-0.5781(0.0298)$.

Following the double pair design (Evans, 1986a), we also considered the following truncated sampling scheme:

$$\pi(\boldsymbol{y}_i) = \begin{cases} \pi_1 = 0 & \text{if } t_i = 0 \\ \pi_2 = 1 & \text{if } t_i > 0 \end{cases} \tag{4.4}$$

This design is similar to (4.2) except that it selects only pairs with at least one death (see Table A.4). Table A.5 provides a comparison of the estimates of α_1 between naive GEE, our estimating function, and the between-within decomposition technique. Once again, naive GEE produces biased estimates. The estimate obtained from our estimating function is unchanged relative to the previous design. This is expected behavior in light of (2.28) which shows that pairs with no deaths are not used. This implies that the efficiency ratio between the estimate based on all pairs and the estimate from our estimating function using only 3204 pairs is still 3.5. This is an improvement over the previous design because we only needed 3204 pairs instead of 6036 pairs to achieve the same efficiency as a random sample of approximately 81000 pairs.

For the truncated sample, the between-within covariate decomposition method produces an estimate that is very different from the untruncated sample: $-0.3525(0.0118)$ versus $-0.5781(0.0298)$.

4.3 Helmet Effectiveness

Evans and Frick (1988) applied the double pair design to assess helmet effectiveness in motorcycle accidents as measured by relative risk of death between riders who wear helmets and riders who do not wear helmets. Because the double pair analysis can accommodate only a single binary covariate, its usage is not feasible if adjustments for other covariates are needed. Greenland (1994) overcame this deficiency by adopting a multiplicative risk modeling approach. He constructed an estimating equation which has direct parallels to conditional likelihood (Breslow and Day, 1980) and generalized Mantel-Haenszel estimating equations (Liang, 1987). For the i-th pair, the multiplicative risk model is described by

$$\log \mu_{ij} = \alpha_i + \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} \tag{4.5}$$

where $\mu_{ij} = \Pr(Y_{ij} = 1|\boldsymbol{x}_{ij})$. If we let y_{1i} and y_{0i} denote the outcomes of the driver and passenger respectively and if $\boldsymbol{z}_i^\top = (\boldsymbol{x}_{1i} - \boldsymbol{x}_{0i})^\top$ denotes the difference in the covariates between driver and passenger, then Greenland's estimating function has the form

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{K} w(\phi_i)\boldsymbol{z}_i\Big(y_{1i} - \phi_i y_{0i}\Big) \tag{4.6}$$

where $\phi_i = \exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})$. The optimal weight function $w(\phi_i)$ depends on the nuisance parameter α_i. To avoid dealing with nuisance parameters, Greenland suggested that reasonable nuisance-free weight functions should be chosen proportional to $1/\phi_i$. He noted that Davis (1985) chose $w(\phi_i) = 1/(\phi_i + 1)$ for odds ratio estimation and suggested that this same weight function can be used for (4.6). This leads to the following unbiased estimating function:

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{K} \boldsymbol{z}_i \left(\frac{y_{1i} - \phi_i y_{0i}}{\phi_i + 1}\right) . \tag{4.7}$$

In sections 3.7.1 and 3.8, we saw that the estimating function (2.28) can be used for multi-

plicative risk models under the sampling scheme

$$\pi(\boldsymbol{y}_i) = \begin{cases} 0, & \text{if } t_i = 0 \\ \pi_2 & \text{if } t_i > 0 \end{cases}$$

The accident data examined by Evans and Frick (1988) and Greenland (1994) falls within the framework of this sampling scheme. The truncated data is provided in table 4.1.

Table 4.1: Motorcycle accident study as reported in Evans and Frick (1988). Male if gender = 0 and female if gender = 1. Wears helmet if helmet = 1 and does not wear helmet if helmet = 0

Driver, Passenger Values:		No. of pairs with death of		
Helmet	Gender	Driver	Passenger	Both
(1, 0)	(0, 0)	70	84	37
(0, 0)	(0, 0)	546	378	226
(1, 0)	(0, 1)	27	36	10
(0, 0)	(0, 1)	342	413	171
(1, 1)	(0, 0)	360	259	152
(0, 1)	(0, 0)	34	8	7
(1, 1)	(0, 1)	279	270	159
(0, 1)	(0, 1)	39	33	6

As is done in Greenland's article, we fit the following multiplicative risk models using our estimating function:

$$\text{Model } 1 : 1 + \text{Driver} + \text{Helmet} + \text{Gender}$$

$$\text{Model } 2 : 1 + \text{Driver} + \text{Helmet} + \text{Gender} + \text{Helmet.Gender}$$

Driver is a binary variable indicating whether the rider is the driver (1) or passenger (0). Helmet is also a binary indicator with value 1 if the rider wears a helmet and 0 otherwise. Table 4.2 compares estimates between our estimating function and Greenland's. Note that for both models, our estimates are essentially identical to Greenland's.

Table 4.2: Comparison of parameter estimates between Greenland's estimating function and ours

	Greenland's	By and Qaqish
Model 1		
Driver	0.241(0.033)	0.242(0.033)
Helmet	-0.340(0.082)	-0.340(0.083)
Gender	0.303(0.049)	0.303(0.049)
Model 2		
Driver	0.240(0.033)	0.240(0.033)
Helmet	-0.317(0.083)	-0.317(0.083)
Gender	-0.354(0.059)	-0.354(0.059)
Helmet.Gender	-0.118(0.072)	-0.118(0.072)

We investigated the reason for this similarity by writing our estimating function in a form similar to Greenland's. It can be shown that for cluster size 2, expression (4.7) is a special case of (2.28).

The advantage of our estimating function over Greenland's is that it was developed under general cluster size in a principled manner using results from the theory of estimating functions. Hence, it can be used in a straightforward manner for any cluster size. On the other hand, Greenland's estimating function, although clever, was obtained in an ad hoc manner by drawing analogies with methods developed for odds ratio estimation. He provided an extension of (4.7) for larger cluster sizes (Greenland, 1994, see expression (13)) but is unsure of its validity.

4.4 Carolina Breast Cancer Study

Overview

The Carolina Breast Cancer Study (CBCS) is a case-control study that has provided some understanding of the risk of breast cancer with respect to environmental exposures and genetic risk factors (Newman et al., 1995) as well as suggesting that breast cancer can be classified into finer subtypes whose distribution depends in an important way on demographic factors (Carey et al., 2006; O'Brien et al., 2010).

One important goal of CBCS was to investigate whether family history of breast cancer is associated with the risk of breast cancer. Towards this end, probands (study participants in the case-control sample) were asked about breast cancer status of relatives (mothers, sisters, and daughters). If at least one relative has breast cancer, then the proband is said to have a history of breast cancer in the family. Our interest with the CBCS data is that it potentially falls within our biased sampling framework if we view the proband and her relatives as the cluster or sampling unit. We may think of breast cancer status within the cluster as a vector of correlated outcomes. Unfortunately, the only covariate common to probands and their relatives is age. Other covariates measured among probands were not measured among relatives. In this sense, we can only consider the following marginal regression model with age as a covariate:

$$\log \mu_{ij} = \alpha + \text{age}_{ij}\, \beta + \boldsymbol{w}_i^\top \boldsymbol{\gamma}. \tag{4.8}$$

where $\mu_{ij} = \Pr(Y_{ij} = 1 \mid \text{age}_{ij})$, $Y_{ij} = 1$ if the j-th member of the i-th cluster has breast cancer and 0 otherwise, and $\boldsymbol{w}_i$ is a vector of unspecified cluster-level covariates.

CBCS Data & Sampling Scheme Considerations

We use the data from phase I (1993-1996) and II (1996-2001) of CBCS. The clustered unit of analysis is the sibship which consists of the proband and her sisters. Our estimation procedure excludes all sibships of size 1 (probands with no sisters). Among sisters of probands, those who are less than 20 years old were excluded from analysis. This is done for the following reasons. First, the small values of age among sisters of control probands is due in part to the fact that the sisters have died in some distant past. It is unreasonable to include these small-valued ages in the analysis. Second, a comparison of the age distribution between cases and controls suggests no subjects fall below 20 years of age.

We can classify the sibships into two types: case sibships, control sibships. Case sibships are sibships made up of the cases (from the case-control study) and her sisters. These sibships have at least one breast cancer member($t_i > 0$). Control sibships are made up of the controls (from the case-control study) and her sisters and these can be further sub-classified as those

with $t_i = 0$ and those with $t_i > 0$ (at least one of the control's sister has breast cancer). For consistency, we assume that control sibships with $t_i > 0$ are actually case sibships in which the breast cancer of the sister was incident prior to Phase I of the study (i.e., prior to 1993). The sample can then be thought of as arising out of the following sampling scheme:

$$\pi(\boldsymbol{y}_i) = \begin{cases} \pi_1(\boldsymbol{d}_i, \text{phase}) & \text{if } t_i = 0 \\ \pi_2(\boldsymbol{d}_i, \text{phase}) & \text{if } t_i > 0 \end{cases}$$

The notation $\pi_j(\boldsymbol{d}_i, \text{phase})$, $j = 1, 2$, indicates that the sampling intensities depend on the phase of the study and on demographic variables $\boldsymbol{d}_i$ which contains age and race/ethnicity of the proband.

Analyses

As mentioned earlier, all sibships of size 1 are excluded from analyses. Our estimation procedure also regards control sibships with $t_i = 0$ as uninformative and hence excludes them from analyses; this is not the same as not sampling families with no diseased members ($\pi_1 = 0$). Our estimating function happens to exclude them. All observations with unknown age are also excluded.

Using our estimating function, we estimate β to be 0.0039(0.004) suggesting that age is not associated with breast cancer given that unspecified cluster-level covariates ($\boldsymbol{w}_i$) are in the model.

Since (4.8) is a marginal model, it is interesting to compare our estimate to those obtain by fitting a logistic regression model with age as the only covariate to the case-control sample only (no siblings or relatives). For this model, we obtained the following estimate of β: 0.0497(0.0037). This value is more than 10 times bigger than the previously estimated value. If we are to believe this estimate, the risk of breast cancer increases by 64% for every 10 years increase in age. Using our estimate, the risk of breast cancer increases by 4% for every 10 years increase in age.

We do not know which numbers are correct since we do not know the true model. The

main difference between our estimate and that from logistic regression is the fact that our approach automatically adjusts for cluster-level covariates where the estimate obtained from logistic regression assumes that the true model contains age as the only covariate, which is probably not realistic.

Chapter 5

Future Research

5.1 Introduction

This dissertation examines two topics: cluster-specific nuisance parameters and outcome-dependent sampling based on cluster totals, the latter being only relevant to correlated binary data. On the surface, the two topics seem very different, but they are in fact related in the sense that this biased sampling problem can be viewed as a nuisance parameter problem. This enabled us to use estimating equations developed for nuisance parameters to make inferences about slope parameters under outcome-dependent sampling based on totals.

While the project has given us greater insights into these matters, it also raises additional issues to which we have yet to provide any answers. We discuss these and other ideas in the following passages with the view that they will serve as a foundation for future research.

5.2 Between-Within Decomposition

Recall from section 2.1 of chapter 2 that the between-within decomposition technique (Neuhaus and McCulloch, 2006) was proposed for fitting GLMMs when the mixing distribution was correlated with covariates in the linear predictor. This violates one of the assumptions of GLMMs. Consider a GLMM with the following linear predictor:

$$\eta_{ij} = b_i + \beta_0 + \beta_1 \, x_{ij} \tag{5.1}$$

To avoid bias under this violation, Neuhaus and McCulloch proposed the between-within decomposition

$$\eta_{ij} = b_i + \beta_0 + \beta_B\ \bar{x}_i + \beta_W\ (x_{ij} - \bar{x}_i) \tag{5.2}$$

They suggested that in-lieu of fitting (5.1), fit a GLMM based on (5.2) as though all GLMM assumptions are satisfied. Their simulation study suggested that $\hat{\beta}_W$ is consistent for β_1 while all other parameter estimates are biased.

There are two conceptual difficulties associated with this approach. First, does β_W have the same interpretation as β_1? If we let z_{ij} denote $(x_{ij} - \bar{x}_i)$, one can argue that it can be interpreted as a contrast in the linear predictor for every unit increase in z_{ij} holding fixed other covariates – including $\bar{x}_i$. But a change in the value of z_{ij} implies a change in the value of $\bar{x}_i$ rendering this interpretation tenuous. A similar difficulty exists with regards to β_B.

In a preliminary investigation, Qaqish (2010, private communication) suggested that in the simple linear regression scenario where all observations within clusters are assumed independent, both $\hat{\beta}_W$ and $\hat{\beta}_B$ estimate β_1. However, this has yet to be examined for situations where the outcomes are correlated and for link functions other than the identity. It would be a worthwhile endeavor to provide clarity with respect to this issue.

5.3 Comparisons With Other Methods

Chapter 2 develops an estimating equations approach for addressing cluster-specific nuisance parameters in regression models. In reviewing the literature, we discussed several other estimating equations approaches that also deal with this problem: conditional GEE (Goetgeluk and Vansteelandt, 2008), projected estimating equations (Rathouz and Liang, 1999), and profile estimating equations (Wang and Hanfelt, 2003). These methods are comparatively much more difficult to use relative to our procedure which can be fit using standard software. In some cases, we showed through a series of propositions (see chapter 2, section 2.11) that our procedure coincides with the method of projection and the method of profiling. Where our method diverges from these other methods (i.e., correlated binary data with log link function), we believe it is worthwhile to compare the performance of these methods with our procedure

using consistency and efficiency as a yardstick. This should give us an idea which method performs better.

5.4 Biased Sampling and Logit-Normal GLMMs

Chapter 3 discusses outcome-dependent sampling based on the total through a biased sampling mechanism that we called variable probability sampling (VPS) (Wooldridge, 1999). Under the setup of GLMMs, we can think of VPS as inducing a GLMM whose mixing distribution is misspecified (Neuhaus and McCulloch, 2006). For the log-normal GLMM, we saw that our estimation procedure is useful for estimating the regression coefficients under VPS based on the total whereas naive estimation leads to inconsistent estimates. However our simulation study for logit-normal GLMMs suggests that naive estimation produces consistent estimates. We have no explanation for why this is the case. Furthermore, we do not know whether this phenomenon is observable for simple models or whether it is observable for all logit-normal GLMMs. I think the reason for this is worth investigating. If we can explain this for all logit-normal GLMMs, then we will have provided an analogue for the results of (Prentice and Pyke, 1979) in the logistic regression setting for GLMMs.

Appendix A

A.1 Evans' Estimator

Let $\boldsymbol{\beta} = [\theta\ \gamma]^\top$, $\boldsymbol{x}_{ij} = [d_{ij}\ s_{ij}]^\top$, and $\zeta_{ij} = \exp\{\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}\}$.

Subtable 1: Belted Driver and Unbelted Passenger Since $d_{i1} = 1$, $d_{i2} = 0$, $s_{i1} = 1$ and $s_{i2} = 0$ we have

$$\boldsymbol{x}_{i1} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{x}_{i2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \sum_{j=1}^{2} \zeta_{ij} = 1 + e^{\theta+\gamma} \quad .$$

This implies that

$$\bar{\boldsymbol{x}}_i = \frac{1}{1+e^{\theta+\gamma}} \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{\theta+\gamma} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} 1 \right) = \begin{bmatrix} \frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \\ \frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \end{bmatrix}$$

so that

$$\boldsymbol{x}_{i1} - \bar{\boldsymbol{x}}_i = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} \frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \\ \frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \end{bmatrix} = \begin{bmatrix} \frac{1}{1+e^{\theta+\gamma}} \\ \frac{1}{1+e^{\theta+\gamma}} \end{bmatrix},$$

$$\boldsymbol{x}_{i2} - \bar{\boldsymbol{x}}_i = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \\ \frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \end{bmatrix} = \begin{bmatrix} -\frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \\ -\frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \end{bmatrix}.$$

Then

$$U_i(\boldsymbol{\beta}) = \begin{bmatrix} \frac{1}{1+e^{\theta+\gamma}} \\ \frac{1}{1+e^{\theta+\gamma}} \end{bmatrix} y_{i1} - \begin{bmatrix} \frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \\ \frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \end{bmatrix} y_{i2} \ . \tag{A.1}$$

Each vehicle in subtable 1 contributes $U_i(\boldsymbol{\beta})$ in (A.1) to $U(\boldsymbol{\beta})$. If we add up all the contributions from subtable 1 we obtain

$$a \begin{bmatrix} \frac{1}{1+e^{\theta+\gamma}} \\ \\ \frac{1}{1+e^{\theta+\gamma}} \end{bmatrix} - b \begin{bmatrix} \frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \\ \\ \frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \end{bmatrix} + c \begin{bmatrix} \frac{1-e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \\ \\ \frac{1-e^{\theta+\gamma}}{1+e^{\theta+\gamma}} \end{bmatrix} \quad . \tag{A.2}$$

Subtable 2: Unbelted Driver and Unbelted Passenger In this subtable $d_{i1} = 1$, $d_{i2} = 0$, $s_{i1} = 0$ and $s_{i2} = 0$ so that

$$\boldsymbol{x}_{i1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{x}_{i2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \sum_{j=1}^{2} \zeta_{ij} = 1 + e^{\theta} \ .$$

This produces

$$\bar{\boldsymbol{x}}_i = \frac{1}{1+e^{\theta}} \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{\theta} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} 1 \right) = \begin{bmatrix} \frac{e^{\theta}}{1+e^{\theta}} \\ 0 \end{bmatrix} \quad .$$

After centering, we have

$$\boldsymbol{x}_{i1} - \bar{\boldsymbol{x}}_i = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{e^{\theta}}{1+e^{\theta}} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{1+e^{\theta}} \\ 0 \end{bmatrix} \ ,$$

$$\boldsymbol{x}_{i2} - \bar{\boldsymbol{x}}_i = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} \frac{e^{\theta}}{1+e^{\theta}} \\ 0 \end{bmatrix} = \begin{bmatrix} -\frac{e^{\theta}}{1+e^{\theta}} \\ 0 \end{bmatrix} \quad .$$

Thus, for all vehicles in this subtable,

$$U_i(\boldsymbol{\beta}) = \begin{bmatrix} \frac{1}{1+e^{\theta}} \\ 0 \end{bmatrix} y_{i1} + \begin{bmatrix} -\frac{e^{\theta}}{1+e^{\theta}} \\ 0 \end{bmatrix} y_{i2} \ . \tag{A.3}$$

Summing over all clusters in subtable 2, we obtain

$$j \begin{bmatrix} \frac{1}{1+e^{\theta}} \\ \\ 0 \end{bmatrix} - k \begin{bmatrix} \frac{e^{\theta}}{1+e^{\theta}} \\ \\ 0 \end{bmatrix} + \ell \begin{bmatrix} \frac{1-e^{\theta}}{1+e^{\theta}} \\ \\ 0 \end{bmatrix} \tag{A.4}$$

The estimating function associated with (2.28) is the sum of (A.2) and (A.4). This gives the estimating equations

$$j\frac{1}{1+e^{\theta}} - k\frac{e^{\theta}}{1+e^{\theta}} + \ell\frac{1-e^{\theta}}{1+e^{\theta}} + a\frac{1}{1+e^{\theta+\gamma}} - b\frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} + c\frac{1-e^{\theta+\gamma}}{1+e^{\theta+\gamma}} = 0$$

$$a\frac{1}{1+e^{\theta+\gamma}} - b\frac{e^{\theta+\gamma}}{1+e^{\theta+\gamma}} + c\frac{1-e^{\theta+\gamma}}{1+e^{\theta+\gamma}} = 0$$

Solving leads to

$$e^{\hat{\theta}} = \frac{j+\ell}{k+\ell} = \frac{m}{n}\ ,$$

$$e^{\hat{\gamma}} = \frac{a+c}{b+c}\cdot\frac{j+\ell}{k+\ell} = \frac{nd}{me}$$

where the second quantity is the Evans' estimator.

A.2 Twins Analyses

Table A.1: Twin pairs data from 1995 to 2000 matched multiple birth file

	Second of twins		
First of twins	Survived 1st year	Died in 1st year	
Survived 1st year	283154	1608	
Died in 1st year	1458	138	
Total			286358

Table A.2: Sample of twin pairs data obtained from biased sampling scheme (4.2). A random sample of size 6036 pairs from the 286358 pairs contains, on average, 68 pairs with at least one death.

	Second of twins		
First of twins	Survived 1st year	Died in 1st year	
Survived 1st year	2832	1608	
Died in 1st year	1458	138	
Total			6036

Table A.3: Parameter estimates of data obtained through biased sampling scheme (4.2)

Parameter	GEE	By and Qaqish	Between-Within†
α_0	0.1628(0.0621)		−0.1054(0.0705)
α_1	−0.2850(0.0132)	−0.4798(0.0382)	−0.5781(0.0298)

†: Estimate of variance component σ_δ^2 is approximately 0

Table A.4: Sample of twin pairs data obtained from biased sampling scheme (4.4). A random sample of size 3204 pairs from the 286358 pairs contains, on average, 36 pairs with at least one death.

	Second of twins		
First of twins	Survived 1st year	Died in 1st year	
Survived 1st year	0	1608	
Died in 1st year	1458	138	
Total			3204

A.3 Conditional Likelihood

Details on the conditional likelihood for independent data are given here for linear regression and Poisson regression. We assume that $Y_1, \ldots, Y_n$ is a random sample of size n.

Table A.5: Parameter estimates of data obtained through biased sampling scheme (4.4)

Parameter	GEE	By and Qaqish	Between-Within†
α_0	$-0.2265(0.0374)$		$-0.5656(0.0499)$
α_1	$-0.0856(0.0077)$	$-0.4798(0.0382)$	$-0.3525(0.0118)$

†: Estimate of variance component σ_δ^2 is approximately 0

Linear Regression

We know that

$$T \sim N\left(n\alpha + \sum_{i=1}^{n} \boldsymbol{x}_i^\top \boldsymbol{\beta}, n\sigma^2\right)$$

Conditional on $T = t$,

$$[\boldsymbol{Y}|T=t] \propto \left(\frac{1}{\sigma^2}\right)^{(n-1)/2} \exp\left\{-\frac{\sum_{i=1}^n (y_i-\mu_i)^2}{2\sigma^2} + \frac{\left(t - n\alpha - \sum_{i=1}^n \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2}{2n\sigma^2}\right\}$$

This implies that the conditional log-likelihood is

$$\ell_c(\boldsymbol{\beta}) \propto \frac{n-1}{2}\log\sigma^2 - \frac{\sum_{i=1}^n (y_i-\mu_i)^2}{2\sigma^2} + \frac{\left(t - n\alpha - \sum_{i=1}^n \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2}{2n\sigma^2}$$

The conditional score function for $\boldsymbol{\beta}$ is

$$\boldsymbol{u}_c(\boldsymbol{\beta}) := \partial_{\boldsymbol{\beta}}\ell_c(\boldsymbol{\beta}) = -\frac{1}{\sigma^2}\sum_{i=1}^n \boldsymbol{x}_i(y_i - \beta_0 - \boldsymbol{x}_i^\top\boldsymbol{\beta}) + \frac{\boldsymbol{X}^\top\mathbf{1}}{n\sigma^2}\left(t - n\alpha - \sum_{i=1}^n \boldsymbol{x}_i^\top\boldsymbol{\beta}\right) .$$

Some simplifications give the score equation

$$\boldsymbol{u}_c(\boldsymbol{\beta}) = \sum_{i=1}^n (\boldsymbol{x}_i^\top - \bar{\boldsymbol{x}})(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}) = \mathbf{0} .$$

where

$$\bar{\boldsymbol{x}} = \frac{\boldsymbol{X}^\top\mathbf{1}}{n} .$$

Poisson Regression

The sufficient statistics for β_0 is T where $T \sim \text{Poisson}(\mathbf{1}^\top \boldsymbol{\mu})$ and

$$\boldsymbol{\mu} = [\mu_1, \ldots, \boldsymbol{\mu}_n]^\top \ .$$

Conditional on $T = t$, the joint distribution of $\boldsymbol{Y}$ is multinomial:

$$\Pr(\boldsymbol{Y} = \boldsymbol{y} | T = t) = \prod_{i=1}^{n} \binom{n}{y_1 \cdots y_n} \pi_i^{y_i} \quad \text{where} \quad \pi_i = \frac{\mu_i}{\mathbf{1}^\top \boldsymbol{\mu}} \ .$$

Thus, the conditional log-likelihood given $T = t$ is

$$\begin{aligned} \ell_c(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i(\beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}) - (\mathbf{1}^\top \boldsymbol{y})\beta_0 - \mathbf{1}^\top \boldsymbol{y} \log(\mathbf{1}^\top \boldsymbol{\zeta}) \\ &= \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} - (\mathbf{1}^\top \boldsymbol{y}) \log(\mathbf{1}^\top \boldsymbol{\zeta}) \end{aligned}$$

where

$$\boldsymbol{\zeta} = (\exp[\boldsymbol{x}_1^\top \boldsymbol{\beta}], \ldots, \exp[\boldsymbol{x}_n^\top \boldsymbol{\beta}])^\top \ .$$

Taking the derivative of ℓ_c with respect to $\boldsymbol{\beta}$ and simplifying gives the conditional score

$$\boldsymbol{u}(\boldsymbol{\beta}) = (\boldsymbol{X} - \bar{x})^\top \boldsymbol{y}$$

where

$$\bar{\boldsymbol{x}} = \frac{\boldsymbol{X}^\top \boldsymbol{\zeta}}{\mathbf{1}^\top \boldsymbol{\zeta}} \ .$$

A.4 Proof Of Propositions

Proposition 1

Before engaging the proof, we set some notations. Write $\boldsymbol{U}_0$ as

$$\boldsymbol{U}_0 = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left[\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}} \right]^\top v_{ij}^{-1}(y_{ij} - \mu_{ij}) = \sum_{i=1}^{K} \boldsymbol{S}_i$$

where

$$\boldsymbol{S}_i = \sum_{j=1}^{n_i} \left[\frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}}\right]^\top v_{ij}^{-1}(y_{ij} - \mu_{ij})$$

Denote by T_{1i} the optimal estimating function for estimating δ_i when $\boldsymbol{\beta}$ is fixed:

$$T_{1i} = \sum_{j=1}^{n_i} \left[\frac{\partial \mu_{ij}}{\partial \delta_i}\right]^\top v_{ij}^{-1}(y_{ij} - \mu_{ij})$$

and let T_{2i} be defined by

$$T_{2i} = \frac{\partial T_{1i}}{\partial \delta_i} + T_{1i}^2 \ .$$

Define $\boldsymbol{T}_i$ by $\boldsymbol{T}_i = [T_{1i},\ T_{2i}]^\top$. The projected estimating function method projects $\boldsymbol{S}_i$ onto the space spanned by $\boldsymbol{T}_i$ (Rathouz and Liang, 1999, last paragraph, page 858).

The notation $\partial_{\delta_i} T_{1i}$ is used interchangeably with $\partial T_{1i}/\partial \delta_i$. The quantities $\boldsymbol{D}_{01}$, D_{11}, D_{12}, D_{21}, $\boldsymbol{D}_{02}$, K_{11}, $D_{22} + 4K_{11}$ are defined in Rathouz and Liang (1999, see Lemmas 2 and 3).

We use r_{ij} to mean the residual $y_{ij} - \mu_{ij}$ and $\boldsymbol{r}_i$ to mean the vector of residuals in the i-th stratum.

The expression $\boldsymbol{A}_2$ is the same quantity as $\boldsymbol{a}_2$ in Rathouz and Liang (1999, expression (7)).

Proof Of Proposition 1 Under Identity Link. Under the canonical identity link, we have

$$\boldsymbol{S}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}(y_{ij} - \mu_{ij}) = \boldsymbol{X}_i^\top \boldsymbol{r}_i$$
$$T_{1i} = \sum_{j=1}^{n_i} y_{ij} - \mu_{ij} = \mathbf{1}^\top \boldsymbol{r}_i \ .$$

Since $\partial_{\delta_i} T_{1i} = -n_i$, it follows that

$$T_{2i} = -n_i + T_{1i}^2 \ .$$

Straightforward calculations show that $\boldsymbol{D}_{01} = \boldsymbol{X}_i^\top \mathbf{1}$, $D_{11} = n_i$, $D_{21} = D_{12} = 0$, $\boldsymbol{D}_{02} = \mathbf{0}$, and $D_{22} + 4K_{11} = 2n_i + 4K_{11}$ where $K_{11} = \binom{n_i}{2}$.

Then

$$\boldsymbol{A}_2 = C(\boldsymbol{\beta}; \delta_i) \begin{bmatrix} \boldsymbol{D}_{01} & \mathbf{0} \end{bmatrix} \begin{bmatrix} D_{22} + 4K_{11} & -D_{12} \\ -D_{21} & D_{11} \end{bmatrix}$$

where

$$C(\boldsymbol{\beta}; \delta_i) = \frac{1}{D_{11}(D_{22} + 4K_{11}) - D_{12}D_{21}} = \frac{1}{D_{11}(D_{22} + 4K_{11})} \ .$$

The projection of $\boldsymbol{S}_i$ onto $\boldsymbol{T}_i$ is

$$\begin{aligned} \boldsymbol{A}_2\boldsymbol{T}_i &= C(\boldsymbol{\beta}; \delta_i) \Big[\boldsymbol{D}_{01}(D_{22} + 4K_{11})\ T_1 \Big] \\ &= \left(\frac{\boldsymbol{D}_{01}}{D_{11}} \right) \mathbf{1}^\top \boldsymbol{r}_i = \left(\frac{\boldsymbol{X}_i^\top \mathbf{1}}{n_i} \right) \mathbf{1}^\top \boldsymbol{r}_i \\ &= \bar{\boldsymbol{x}}_i\ (\mathbf{1}^\top \boldsymbol{r}_i) \end{aligned}$$

where $\bar{\boldsymbol{x}}_i$ is the simple mean of $\boldsymbol{X}_i$. Then the corrected estimating function $\boldsymbol{S}_{2i}$ is obtained as follows:

$$\begin{aligned} \boldsymbol{S}_{2i} = \boldsymbol{S}_i - \boldsymbol{A}_2\boldsymbol{T}_i &= \boldsymbol{X}_i^\top \boldsymbol{r}_i - \bar{\boldsymbol{x}}_i\ (\mathbf{1}^\top \boldsymbol{r}_i) \\ &= (\boldsymbol{X}_i^\top - \bar{\boldsymbol{x}}_i\ \mathbf{1}^\top)(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\ &= (\boldsymbol{X}_i^\top - \bar{\boldsymbol{x}}_i\ \mathbf{1}^\top)(\boldsymbol{y}_i - \mathbf{1}\delta_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\ &= (\boldsymbol{X}_i - \mathbf{1}\bar{\boldsymbol{x}}_i^\top)^\top (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \end{aligned}$$

where the last equality follows from the fact that $(\boldsymbol{X}_i - \mathbf{1}\bar{\boldsymbol{x}}_i^\top)^\top \mathbf{1}\delta_i = \mathbf{0}$. This gives the projected estimating function

$$\boldsymbol{U}_{PEF} = \sum_{i=1}^{K} \boldsymbol{S}_{2i} = \sum_{i=1}^{K} (\boldsymbol{X}_i - \mathbf{1}\bar{\boldsymbol{x}}_i^\top)^\top (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

which is equal to (2.25). □

Proof of Proposition 1 Under Log Link. We have

$$\begin{aligned}\boldsymbol{S}_i &= \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}(y_{ij} - \mu_{ij}) = \boldsymbol{X}_i^\top \boldsymbol{r}_i \\ T_{1i} &= \sum_{j=1}^{n_i} y_{ij} - \mu_{ij} = \mathbf{1}^\top \boldsymbol{r}_i \ .\end{aligned}$$

Since $\partial_{\delta_i} T_{1i} = -\sum_{j=1}^{n_i} \mu_{ij}$, it follows that

$$T_{2i} = -\sum_{j=1}^{n_i} \mu_{ij} + T_{1i}^2 \ .$$

Direct calculation leads to

$$\begin{aligned}\boldsymbol{D}_{01} &= \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}\mu_{ij} = \boldsymbol{X}_i^\top \boldsymbol{\mu}_i = \boldsymbol{D}_{02} \ , \\ D_{11} = D_{21} = D_{12} &= \sum_{j=1}^{n_i} \mu_{ij} = \mathbf{1}^\top \boldsymbol{\mu}_i \ , \\ D_{22} + 4K_{11} &= \sum_{j=1}^{n_i} \Big\{ 2\mu_{ij}^2 + \mu_{ij} \Big\} + 4K_{11} \ , \\ K_{11} &= \sum_{j=1}^{n_i - 1} \sum_{\ell=j+1}^{n_i} \mu_{ij}\mu_{i\ell} \ .\end{aligned}$$

Therefore, $\boldsymbol{A}_2$ is equal to

$$\boldsymbol{A}_2 = C(\boldsymbol{\beta}; \delta_i) \begin{bmatrix} \boldsymbol{D}_{01} & \boldsymbol{D}_{02} \end{bmatrix} \begin{bmatrix} D_{22} + 4K_{11} & -D_{12} \\ -D_{21} & D_{11} \end{bmatrix}$$

where

$$C(\boldsymbol{\beta}; \delta_i) = \frac{1}{D_{11}(D_{22} + 4K_{11}) - D_{21}D_{12}} \ .$$

Since $D_{11} = D_{21} = D_{12}$ it follows that

$$
\begin{aligned}
\boldsymbol{A}_2\boldsymbol{T}_i &= C(\boldsymbol{\beta};\delta_i)\Big\{\Big[\boldsymbol{D}_{01}(D_{22}+4K_{11})-\boldsymbol{D}_{02}D_{21}\Big]T_{1i}+\Big[\boldsymbol{D}_{02}D_{11}-\boldsymbol{D}_{01}D_{12}\Big]T_{2i}\Big\}\\
&= C(\boldsymbol{\beta};\delta_i)\Big\{\Big[\boldsymbol{D}_{01}(D_{22}+4K_{11})-\boldsymbol{D}_{02}D_{21}\Big]T_{1i}\Big\}\\
&= \frac{\Big[\boldsymbol{D}_{01}\big(D_{22}+4K_{11}-D_{11}\big)\mathbf{1}^\top\boldsymbol{r}_i\Big]}{D_{11}\big(D_{22}+4K_{11}-D_{11}\big)}\\
&= \left(\frac{\boldsymbol{D}_{01}}{D_{11}}\right)\mathbf{1}^\top\boldsymbol{r}_i = \left(\frac{\boldsymbol{X}_i^\top\boldsymbol{\mu}_i}{\mathbf{1}^\top\boldsymbol{\mu}_i}\right)\mathbf{1}^\top\boldsymbol{r}_i\\
&= \left(\frac{\boldsymbol{X}_i^\top\boldsymbol{\zeta}_i}{\mathbf{1}^\top\boldsymbol{\zeta}_i}\right)\mathbf{1}^\top\boldsymbol{r}_i\\
&= (\bar{\boldsymbol{x}}_i\mathbf{1}^\top)\boldsymbol{r}_i
\end{aligned}
$$

where $\bar{\boldsymbol{x}}_i$ is the mean of $\boldsymbol{X}_i$ weighted by $\boldsymbol{\zeta}_i := \exp[\boldsymbol{X}_i\boldsymbol{\beta}]$. Then

$$
\begin{aligned}
\boldsymbol{S}_{2i} &= \boldsymbol{S}_i - \boldsymbol{A}_2\boldsymbol{T}_i = \boldsymbol{X}_i^\top\boldsymbol{r}_i - \bar{\boldsymbol{x}}_i\mathbf{1}^\top\boldsymbol{r}_i\\
&= \Big(\boldsymbol{X}_i^\top - \bar{\boldsymbol{x}}_i\mathbf{1}^\top\Big)(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\\
&= \Big(\boldsymbol{X}_i^\top - \bar{\boldsymbol{x}}_i\mathbf{1}^\top\Big)\boldsymbol{y}_i\\
&= (\boldsymbol{X}_i - \mathbf{1}\bar{\boldsymbol{x}}_i^\top)^\top\boldsymbol{y}_i\ . \qquad \text{(A.5)}
\end{aligned}
$$

Thus,

$$
\boldsymbol{U}_{PEF} = \sum_{i=1}^{K}\boldsymbol{S}_{2i} = \sum_{i=1}^{K}(\boldsymbol{X}_i - \mathbf{1}\bar{\boldsymbol{x}}_i^\top)^\top\boldsymbol{y}_i
$$

which is equal to (2.28). □

Proposition 2

The proof requires the same notational setup as Proposition 1.

Proof of Proposition 2 Under Gamma Variance. If $v_{ij} = \mu_{ij}^2$, then

$$\boldsymbol{S}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \left(\frac{y_{ij} - \mu_{ij}}{\mu_{ij}} \right) \ , \quad T_{1i} = \sum_{j=1}^{n_i} \left(\frac{y_{ij} - \mu_{ij}}{\mu_{ij}} \right) \ , \quad T_{2i} = \partial_{\delta_i} T_{1i} + T_{1i}^2 \ .$$

Direct calculation leads to

$$\begin{aligned} \boldsymbol{D}_{01} &= \boldsymbol{X}_i^\top \mathbf{1}, \quad D_{11} = \mathbf{1}^\top \mathbf{1} = n_i, \quad D_{21} = \mathbf{1}^\top \mathbf{1} = n_i \\ \boldsymbol{D}_{02} &= \boldsymbol{X}_i^\top \mathbf{1}, \quad D_{12} = \mathbf{1}^\top \mathbf{1} = n_i, \quad D_{22} = 5n_i \end{aligned}$$

Then

$$\boldsymbol{A}_2 = C(\boldsymbol{\beta}, \delta_i) \begin{bmatrix} \boldsymbol{X}_i^\top \mathbf{1}(D_{22} + 4K_{11} - n_i) & \mathbf{0} \end{bmatrix}$$

where

$$C(\boldsymbol{\beta}, \delta_i) = \frac{1}{\mathbf{1}^\top \mathbf{1}(D_{22} + 4K_{11} - n_i)}$$

Then

$$\boldsymbol{A}_2 \boldsymbol{T}_i = \frac{\boldsymbol{X}_i^\top \mathbf{1}}{\mathbf{1}^\top \mathbf{1}} T_{1i} = \bar{\boldsymbol{x}}_i T_{1i}$$

where $\bar{\boldsymbol{x}}_i$ is the simple average. This leads to

$$\boldsymbol{S}_{2i} = \boldsymbol{S}_i - \boldsymbol{A}_2 \boldsymbol{T}_i = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i) \left(\frac{y_{ij} - \mu_{ij}}{\mu_{ij}} \right)$$

Thus,

$$\boldsymbol{U}_{PEF} = \sum_{i=1}^{K} \boldsymbol{S}_{2i} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i) \left(\frac{y_{ij} - \mu_{ij}}{\mu_{ij}} \right)$$

which is equal to (2.33) where $\bar{\boldsymbol{x}}_i$ is defined by the weights $c_{ij}\mu_{ij} = 1$. □

Proof of Proposition 2 Under Inverse Gaussian Variance. If $v_{ij} = \mu_{ij}^3$, then

$$\boldsymbol{S}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} \left(\frac{y_{ij} - \mu_{ij}}{\mu_{ij}^2} \right) \ , \quad T_{1i} = \sum_{j=1}^{n_i} \left(\frac{y_{ij} - \mu_{ij}}{\mu_{ij}^2} \right) \ , \quad T_{2i} = \partial_{\delta_i} T_{1i} + T_{1i}^2 \ .$$

Direct calculation leads to

$$\begin{aligned} \boldsymbol{D}_{01} &= \boldsymbol{X}_i^\top \boldsymbol{\vartheta}_i, \quad D_{11} = \mathbf{1}^\top \boldsymbol{\vartheta}_i, \quad D_{21} = \mathbf{1}^\top \boldsymbol{\vartheta}_i \\ \boldsymbol{D}_{02} &= \boldsymbol{X}_i^\top \boldsymbol{\vartheta}_i, \quad D_{12} = \mathbf{1}^\top \boldsymbol{\vartheta}_i, \quad D_{22} = 7\mathbf{1}^\top \boldsymbol{\vartheta}_i + 2\boldsymbol{\vartheta}_i^\top \boldsymbol{\vartheta}_i \end{aligned}$$

where

$$\boldsymbol{\vartheta}_i = \frac{1}{\boldsymbol{\mu}_i}$$

Then,

$$\boldsymbol{A}_2 = C(\boldsymbol{\beta}, \delta_i) \left[\boldsymbol{X}_i^\top \boldsymbol{\vartheta}_i (D_{22} + 4K_{11} - \mathbf{1}^\top \boldsymbol{\vartheta}_i) \quad \mathbf{0} \right]$$

where

$$C(\boldsymbol{\beta}, \delta_i) = \frac{1}{\mathbf{1}^\top \boldsymbol{\vartheta}_i (D_{22} + 4K_{11} - \mathbf{1}^\top \boldsymbol{\vartheta}_i)}$$

which leads to

$$\boldsymbol{A}_2 \boldsymbol{T}_i = \bar{\boldsymbol{x}}_i T_{1i}$$

where

$$\bar{\boldsymbol{x}}_i = \frac{\boldsymbol{X}_i^\top \boldsymbol{\vartheta}_i}{\mathbf{1}^\top \boldsymbol{\vartheta}_i} \ .$$

Thus,

$$\boldsymbol{S}_{2i} = \boldsymbol{S}_i - \boldsymbol{A}_2 \boldsymbol{T}_i = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i) \left(\frac{y_{ij} - \mu_{ij}}{\mu_{ij}^2} \right)$$

and

$$\boldsymbol{U}_{PEF} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i) \left(\frac{y_{ij} - \mu_{ij}}{\mu_{ij}^2} \right) \ .$$

This is the same as (2.33) with $\bar{\boldsymbol{x}}_i$ defined by the weights $c_{ij}\mu_{ij} = 1/\mu_{ij}$. □

Proposition 3

Before proceeding to the proof, we establish some notations:

$$\boldsymbol{X}_i = \begin{bmatrix} x_{i11} & \cdots & x_{i1p} \\ \vdots & \ddots & \vdots \\ x_{in_i1} & \cdots & x_{in_ip} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_{i1}^\top \\ \vdots \\ \boldsymbol{x}_{in_i}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_{i.1} & \cdots & \boldsymbol{x}_{i.p} \end{bmatrix}$$

where $\boldsymbol{X}_i$ is the design matrix for the i-th cluster, $\boldsymbol{x}_{ij}^\top$ is the j-th row of $\boldsymbol{X}_i$, and $\boldsymbol{x}_{i.k}$ is the k-th column of $\boldsymbol{X}_i$.

When the link is canonical to the variance function, we can write $\boldsymbol{U}_0$ as

$$\begin{aligned} \boldsymbol{U}_0 &:= \sum_{i=1}^{K} \boldsymbol{U}_i(\boldsymbol{\beta}; \delta_i) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}(y_{ij} - \mu_{ij}) \\ &= \begin{bmatrix} U_{01} \\ \vdots \\ U_{0p} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ij1}(y_{ij} - \mu_{ij}) \\ \vdots \\ \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ijp}(y_{ij} - \mu_{ij}) \end{bmatrix} \end{aligned}$$

The goal is to find a first-order ancillary estimating function $\boldsymbol{g}$ from $\boldsymbol{U}_0$ of the form

$$\boldsymbol{g} = \begin{bmatrix} g_1 & \cdots & g_p \end{bmatrix}^\top$$

after which adjust it to obtain $\boldsymbol{g}_{adj}$ that is approximately unbiased. Wang and Hanfelt suggested a procedure that constructs g_1 through g_p separately. Without loss of generality, we construct only g_1 and conclude that g_2,..., g_p are obtained by the same procedure.

We will use the fact that for fixed $\boldsymbol{\beta}$, the optimal estimating function for estimating δ_i is

$$h_i := \sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \delta_i} \left(\frac{y_{ij} - \mu_{ij}}{v_{ij}} \right) .$$

which is equal to

$$h_i := \sum_{j=1}^{n_i} (y_{ij} - \mu_{ij})$$

when the link is canonical to the variance function.

Proof of Proposition 3 Under Log Link. First, the derivative of h_i with respect to δ_i is

$$\partial_{\delta_i} h_i = -\sum_{j=1}^{n_i} \mu_{ij} = \mathbf{1}^\top \boldsymbol{\mu}_i \ .$$

Letting

$$\boldsymbol{h} \equiv \boldsymbol{h}(\boldsymbol{\delta}) = \begin{bmatrix} h_1 \\ \vdots \\ h_K \end{bmatrix}$$

and using results from Magnus and Neudecker (1999) and Wand (2002), we have $\partial_{\boldsymbol{\delta}} \boldsymbol{h} = \text{diag}\Big\{\partial_{\delta_1} h_1, \ldots, \partial_{\delta_K} h_K\Big\}$. Under the log link, the derivative of U_{01} with respect to $\boldsymbol{\delta}$ is

$$\partial_{\boldsymbol{\delta}} U_{01} = -\left[\textstyle\sum_{j=1}^{n_1} x_{1j1}\mu_{1j}, \cdots, \sum_{j=1}^{n_K} x_{Kj1}\mu_{Kj}\right] := -\left[\boldsymbol{x}_{1.1}^\top \boldsymbol{\mu}_1, \cdots, \boldsymbol{x}_{K.1}^\top \boldsymbol{\mu}_K\right]$$

where $\boldsymbol{\mu}_i = [\mu_{i1}, \ldots, \mu_{in_i}]^\top$. Next, construct g_1 by

$$\begin{aligned}
g_1 &= U_{01} - E[\partial_{\boldsymbol{\delta}} U_{01}] E^{-1}[\partial_{\boldsymbol{\delta}} \boldsymbol{h}] \boldsymbol{h} \\
&= U_{01} - \left[\boldsymbol{x}_{1.1}^\top \boldsymbol{\mu}_1, \cdots, \boldsymbol{x}_{K.1}^\top \boldsymbol{\mu}_K\right] \text{diag}\Big\{\frac{1}{\mathbf{1}^\top \boldsymbol{\mu}_1}, \ldots, \frac{1}{\mathbf{1}^\top \boldsymbol{\mu}_K}\Big\} \boldsymbol{h} \\
&= U_{01} - \begin{bmatrix} \frac{\boldsymbol{x}_{1.1}^\top \boldsymbol{\mu}_1}{\mathbf{1}^\top \boldsymbol{\mu}_1} & \cdots & \frac{\boldsymbol{x}_{K.1}^\top \boldsymbol{\mu}_K}{\mathbf{1}^\top \boldsymbol{\mu}_K} \end{bmatrix} \boldsymbol{h} \\
&= U_{01} - \sum_{i=1}^{K} \left(\frac{\boldsymbol{x}_{i.1}^\top \boldsymbol{\mu}_i}{\mathbf{1}^\top \boldsymbol{\mu}_1}\right) h_i := U_{01} - \sum_{i=1}^{K} \bar{x}_{i1} h_i \\
&= \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij1} - \bar{x}_{i1})(y_{ij} - \mu_{ij})
\end{aligned}$$

Note that $\bar{x}_{i1}$ is the weighted average of the first covariate $\boldsymbol{x}_{i.1}$ where the weight is the mean μ_{ij}. Let Q_1 denote the correction term. From Wang and Hanfelt (2003, Theorem 1),

$$Q_1 = -\frac{1}{2}\,\mathrm{trace}\Big\{E^{-1}[\partial_{\boldsymbol{\delta}}\boldsymbol{h}]E[\partial_{\boldsymbol{\delta}\boldsymbol{\delta}^\top} g_1]\Big\}\ .$$

In this expression,

$$\partial_{\boldsymbol{\delta}\boldsymbol{\delta}^\top} g_1 = \frac{\partial^2 g_1}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^\top}\ .$$

Next,

$$\partial_{\delta_i} g_1 = -\sum_{j=1}^{n_i}(x_{ij1} - \bar{x}_{i1})\mu_{ij} - \partial_{\delta_i}\bar{x}_{i1}\sum_{j=1}^{n_i}(y_{ij} - \mu_{ij})$$

and

$$\partial_{\delta_i\delta_i} g_1 = -\sum_{j=1}^{n_i}(x_{ij1} - \bar{x}_{i1})\mu_{ij} - \partial_{\delta_i}\bar{x}_{i1}\mu_{ij} + \partial_{\delta_i}\bar{x}_{i1}\mu_{ij} - \left(\partial_{\delta_i\delta_i}\bar{x}_{i1}\right)(y_{ij} - \mu_{ij})$$

The first summand is zero and the expected value of the last summand is zero. Thus $E\left[\partial_{\delta_i\delta_i} g_1\right] = \mathbf{0}$ so that $Q_1 = 0$. This implies that $g_{1,adj} = g_1$ and therefore

$$\begin{aligned}\boldsymbol{g}_{adj} &= \begin{bmatrix} g_{1,adj} \\ \vdots \\ g_{p,adj}\end{bmatrix} = \begin{bmatrix} \sum_{i=1}^K\sum_{j=1}^{n_i}(x_{ij1} - \bar{x}_{i1})(y_{ij} - \mu_{ij}) \\ \vdots \\ \sum_{i=1}^K\sum_{j=1}^{n_i}(x_{ijp} - \bar{x}_{ip})(y_{ij} - \mu_{ij})\end{bmatrix} \\ &= \sum_{i=1}^K\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(y_{ij} - \mu_{ij}) = \sum_{i=1}^K\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)y_{ij}\end{aligned}$$

where

$$\bar{\boldsymbol{x}}_i = \frac{\boldsymbol{X}_i^\top\boldsymbol{\mu}_i}{\mathbf{1}^\top\boldsymbol{\mu}_i} = \frac{\boldsymbol{X}_i^\top\boldsymbol{\zeta}_i}{\mathbf{1}^\top\boldsymbol{\zeta}_i}\ .$$

This is exactly expression (2.28). □

Proof of Proposition 3 Under Identity Link. Under the canonical identity link, we have

$$\partial_{\delta_i} h_i = -\sum_{j=1}^{n_i} 1 = -n_i$$

and

$$\partial_{\boldsymbol{\delta}} U_{01} = \begin{bmatrix} -\sum_{j=1}^{n_i}\boldsymbol{x}_{1j1} & \cdots & -\sum_{j=1}^{n_i}\boldsymbol{x}_{Kj1}\end{bmatrix} = \begin{bmatrix} -\boldsymbol{x}_{1.1}^\top\mathbf{1} & \cdots & \boldsymbol{x}_{K.1}^\top\mathbf{1}\end{bmatrix}\ .$$

This gives

$$\begin{aligned} g_1 &= U_{01} - E([\partial_{\boldsymbol{\delta}} U_{01}] E^{-1} [\partial_{\boldsymbol{\delta}} \boldsymbol{h}] \boldsymbol{h} \\ &= U_{01} - \left[\boldsymbol{x}_{1.1}^\top \mathbf{1} \quad \cdots \quad \boldsymbol{x}_{K.1}^\top \mathbf{1} \right] \text{diag} \left(\frac{1}{n_1}, \ldots, \frac{1}{n_K} \right) \boldsymbol{h} \\ &= U_{01} - \left[\frac{\boldsymbol{x}_{1.1}^\top \mathbf{1}}{n_1} \quad \ldots \quad \frac{\boldsymbol{x}_{K.1}^\top \mathbf{1}}{n_K} \right] \boldsymbol{h} = U_{01} - \sum_{i=1}^{K} \frac{\boldsymbol{x}_{i.1}^\top \mathbf{1}}{n_i} h_i \\ &= U_{01} - \sum_{i=1}^{K} \sum_{j=1}^{n_i} \bar{x}_{i1} (y_{ij} - \mu_{ij}) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij1} - \bar{x}_{i1})(y_{ij} - \mu_{ij}) \end{aligned}$$

where $\bar{x}_{i1}$ is the simple average of the first column in $\boldsymbol{X}_i$. For simple averages, we know that $\sum_{j=1}^{n_i} (x_{ij1} - \bar{x}_{i1}) = 0$. Using this fact, g_1 can be reduced as follows:

$$\begin{aligned} g_1 &= \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij1} - \bar{x}_{i1})(y_{ij} - \delta_i - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}) \\ &= \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij1} - \bar{x}_{i1})(y_{ij} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}) \end{aligned}$$

To derive the correction term Q_1, first note that g_1 is $\boldsymbol{\delta}$-free. It follows that $E\{\partial_{\boldsymbol{\delta}\boldsymbol{\delta}^\top} g_1\} = \mathbf{0}$ so that $Q_1 = 0$. Thus,

$$\boldsymbol{g}_{adj} = \boldsymbol{g} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(y_{ij} - \boldsymbol{x}_{ij}^\top \boldsymbol{\beta})$$

But this is exactly the covariate-centered estimating functiongiven in (2.25). □

Proposition 4

Proof of Proposition 4 Under Projected Estimating Function. Using the notations given in the proof of Proposition 1, we have

$$\boldsymbol{S}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} (1 + \psi_{ij})(y_{ij} - \mu_{ij}) \ , \quad T_{1i} = h_i = \sum_{j=1}^{n_i} (1 + \psi_{ij})(y_{ij} - \mu_{ij}) \ , \quad T_{2i} = \partial_{\delta_i} T_{1i} + T_{1i}^2$$

and $\boldsymbol{T}_i = [T_{1i} \ \ T_{2i}]^\top$. Straightforward calculation leads to

$$\boldsymbol{D}_{01} = \boldsymbol{X}_i^\top \boldsymbol{\psi}_i, \quad D_{11} = \mathbf{1}^\top \boldsymbol{\psi}_i, \quad D_{21} = \mathbf{1}^\top \boldsymbol{\psi}_i,$$
$$\boldsymbol{D}_{02} = \boldsymbol{X}_i^\top \boldsymbol{\psi}_i, \quad D_{12} = \mathbf{1}^\top \boldsymbol{\psi}_i$$

Then

$$\begin{aligned}\boldsymbol{A}_2 &= C(\boldsymbol{\beta}, \delta_i)[\boldsymbol{D}_{01}(D_{22} + 4K_{11}) - \boldsymbol{D}_{02}D_{21} \quad - \boldsymbol{D}_{01}D_{12} + \boldsymbol{D}_{02}D_{11}] \\ &= C(\boldsymbol{\beta}, \delta_i)[\boldsymbol{X}_i^\top \boldsymbol{\psi}_i(D_{22} + 4K_{11} - \mathbf{1}^\top \boldsymbol{\psi}_i) \quad \mathbf{0}]\end{aligned}$$

Since

$$C(\boldsymbol{\beta}, \delta_i) = \frac{1}{\mathbf{1}^\top \boldsymbol{\psi}_i(D_{22} + 4K_{11} - \mathbf{1}^\top \boldsymbol{\psi}_i)}$$

it follows that $\boldsymbol{A}_2$ reduces to

$$\boldsymbol{A}_2 = [\bar{\boldsymbol{x}}_i \;\; \mathbf{0}] \quad \text{where} \quad \bar{\boldsymbol{x}}_i = \frac{\boldsymbol{X}_i^\top \boldsymbol{\psi}_i}{\mathbf{1}^\top \boldsymbol{\psi}_i}$$

Thus, the projected term is

$$\boldsymbol{A}_2 \boldsymbol{T}_i = \bar{\boldsymbol{x}}_i T_{1i} = \bar{\boldsymbol{x}}_i \sum_{j=1}^{n_i} (1 + \psi_{ij})(y_{ij} - \mu_{ij})$$

This implies that

$$\boldsymbol{S}_{2i} = \boldsymbol{S}_i - \boldsymbol{A}_2 \boldsymbol{T}_i = \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(1 + \psi_{ij})(y_{ij} - \mu_{ij})$$

and that

$$\boldsymbol{U}_{PEF} = \sum_{i=1}^{K} \boldsymbol{S}_i = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(1 + \psi_{ij})(y_{ij} - \mu_{ij})$$

which is equal to (2.44). □

Proof of Proposition 4 Under Adjusted Profile Estimating Function. The notations are the same as those used in the proof of Proposition 3

Write

$$\boldsymbol{U}_0 = \begin{bmatrix} U_{01} \\ \vdots \\ U_{0p} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ij1}(1+\psi_{ij})(y_{ij}-\mu_{ij}) \\ \vdots \\ \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ijp}(1+\psi_{ij})(y_{ij}-\mu_{ij}) \end{bmatrix} .$$

Since

$$\partial_{\delta_i} U_{01} = \sum_{j=1}^{n_i} x_{ij1} \left\{ \frac{-\mu_{ij}(1-\mu_{ij}) + (y_{ij}-\mu_{ij})\mu_{ij}}{(1-\mu_{ij})^2} \right\}$$

it follows that

$$\begin{aligned} E\left[\partial_{\boldsymbol{\delta}} U_{01}\right] &= \Big[-\sum_{j=1}^{n_1} x_{1j1}\psi_{1j} \quad \cdots \quad -\sum_{j=1}^{n_K} x_{Kj1}\psi_{Kj} \Big] \\ &= [-\boldsymbol{x}_{1.1}^{\top}\boldsymbol{\psi}_1 \quad \cdots \quad -\boldsymbol{x}_{K.1}^{\top}\boldsymbol{\psi}_K] \end{aligned}$$

Similarly, since

$$\partial_{\delta_i} h_i = \sum_{j=1}^{n_i} \frac{-\mu_{ij}(1-\mu_{ij}) + (y_{ij}-\mu_{ij})\mu_{ij}}{(1-\mu_{ij})^2}$$

and

$$E\left[\partial_{\delta_i} h_i\right] = -\sum_{j=1}^{n_i} \psi_{ij} = -\mathbf{1}^{\top}\boldsymbol{\psi}_i$$

it follows that

$$E\left[\partial_{\boldsymbol{\delta}} \boldsymbol{h}\right] = \operatorname{diag}\left\{ -\mathbf{1}^{\top}\boldsymbol{\psi}_1, \ldots, \mathbf{1}^{\top}\boldsymbol{\psi}_K \right\} .$$

Then,

$$\begin{aligned}
g_1 &= U_{01} - E\left[\partial_{\boldsymbol{\delta}} U_{01}\right] E^{-1}\left[\partial_{\boldsymbol{\delta}} \boldsymbol{h}\right] \boldsymbol{h} \\
&= U_{01} - \begin{bmatrix} \frac{\boldsymbol{x}_{1.1}^\top \boldsymbol{\psi}_1}{\mathbf{1}^\top \boldsymbol{\psi}_1} & \dots & \frac{\boldsymbol{x}_{K.1}^\top \boldsymbol{\psi}_K}{\mathbf{1}^\top \boldsymbol{\psi}_K} \end{bmatrix} \boldsymbol{h} = U_{01} - \sum_{i=1}^{K} \frac{\boldsymbol{x}_{i.1}^\top \boldsymbol{\psi}_i}{\mathbf{1}^\top \boldsymbol{\psi}_i} h_i \\
&= \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{ij1}(1+\psi_{ij})(y_{ij}-\mu_{ij}) - \sum_{i=1}^{K} \sum_{j=1}^{n_i} \bar{x}_{i1}(1+\psi_{ij})(y_{ij}-\mu_{ij}) \\
&= \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij1}-\bar{x}_{i1})(1+\psi_{ij})(y_{ij}-\mu_{ij})
\end{aligned}$$

where

$$\bar{x}_{i1} = \frac{\boldsymbol{x}_{i.1}^\top \boldsymbol{\psi}_i}{\mathbf{1}^\top \boldsymbol{\psi}_i}$$

is the weighted average of the first column of $\boldsymbol{X}_i$ where the weights are the odds. Then

$$\boldsymbol{g} = \begin{bmatrix} \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ij1}-\bar{x}_{i1})(1+\psi_{ij})(y_{ij}-\mu_{ij}) \\ \vdots \\ \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{ijp}-\bar{x}_{ip})(1+\psi_{ij})(y_{ij}-\mu_{ij}) \end{bmatrix} = \sum_{i=1}^{K} \boldsymbol{X}_{ci}^\top \big\{ (\mathbf{1}+\boldsymbol{\psi}_i) \odot (\boldsymbol{y}_i - \boldsymbol{\mu}_i) \big\}$$

The correction term for g_1 is

$$Q_1 := -\frac{1}{2} \operatorname{trace}\left\{ E^{-1}\left[\partial_{\boldsymbol{\delta}} \boldsymbol{h}\right] E\left[\partial_{\boldsymbol{\delta}\boldsymbol{\delta}^\top} g_1\right] \right\} \ .$$

Next,

$$\begin{aligned}
\partial_{\delta_i} g_1 &= -\sum_{j=1}^{n_i} \partial_{\delta_i} \bar{x}_{i1}(1+\psi_{ij})(y_{ij}-\mu_{ij}) + (x_{ij1}-\bar{x}_{i1}) \left\{ \frac{-\mu_{ij}(1-\mu_{ij}) + (y_{ij}-\mu_{ij})\mu_{ij}}{(1-\mu_{ij})^2} \right\} \\
&= -\sum_{j=1}^{n_i} \partial_{\delta_i} \bar{x}_{i1}(1+\psi_{ij})(y_{ij}-\mu_{ij}) - \sum_{j=1}^{n_i} (x_{ij1}-\bar{x}_{i1})\psi_{ij} \\
&\qquad + \sum_{j=1}^{n_i} (x_{ij1}-\bar{x}_{i1})\psi_{ij}(1+\psi_{ij})(y_{ij}-\mu_{ij}) \\
&= \sum_{j=1}^{n_i} \Big\{ (x_{ij1}-\bar{x}_{i1})\psi_{ij} - \partial_{\delta_i} \bar{x}_{i1} \Big\} \Big\{ (1+\psi_{ij})(y_{ij}-\mu_{ij}) \Big\}
\end{aligned}$$

The last equality follows from the fact that $\sum_{j=1}^{n_i}(x_{ij1}-\bar{x}_{i1})\psi_{ij}=0$. Letting

$$c_{ij}=(x_{ij1}-\bar{x}_{i1})\psi_{ij}-\partial_{\delta_i}\bar{x}_{i1}$$

we have

$$\partial_{\delta_i\delta_i}g_1=\sum_{j=1}^{n_i}\partial_{\delta_i}c_{ij}\Big\{(1+\psi_{ij})(y_{ij}-\mu_{ij})\Big\}+\sum_{j=1}^{n_i}c_{ij}\left\{\frac{-\mu_{ij}(1-\mu_{ij})+(y_{ij}-\mu_{ij})\mu_{ij}}{(1-\mu_{ij})^2}\right\}$$

Taking expectation leads to

$$E\left[\partial_{\delta_i\delta_i}g_1\right]=-\sum_{j=1}^{n_i}\Big\{(x_{ij1}-\bar{x}_{i1})\psi_{ij}-\partial_{\delta_i}\bar{x}_{i1}\Big\}\psi_{ij}$$

so that

$$E\left[\partial_{\boldsymbol{\delta}\boldsymbol{\delta}^\top}g_1\right]=-\operatorname{diag}\left\{\textstyle\sum_{j=1}^{n_i}\Big[(x_{ij1}-\bar{x}_{i1})\psi_{ij}-\partial_{\delta_i}\bar{x}_{i1}\Big]\psi_{ij}\right\}_{i=1}^{K}$$

This implies that

$$\begin{aligned}Q_1&=-\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{\left[(x_{ij1}-\bar{x}_{i1})\psi_{ij}-\partial_{\delta_i}\bar{x}_{i1}\right]\psi_{ij}}{\mathbf{1}^\top\boldsymbol{\psi}_i}\\&=-\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{\left[(x_{ij1}-\bar{x}_{i1})\psi_{ij}^2\right]}{\mathbf{1}^\top\boldsymbol{\psi}_i}+\frac{1}{2}\sum_{i=1}^{K}\partial_{\delta_i}\bar{x}_{i1}\end{aligned}$$

Write

$$\begin{aligned}\boldsymbol{Q}=\begin{bmatrix}Q_1\\\vdots\\Q_p\end{bmatrix}&=\begin{bmatrix}-\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{\left[(x_{ij1}-\bar{x}_{i1})\psi_{ij}^2\right]}{\mathbf{1}^\top\boldsymbol{\psi}_i}+\frac{1}{2}\sum_{i=1}^{K}\partial_{\delta_i}\bar{x}_{i1}\\\vdots\\-\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{\left[(x_{ijp}-\bar{x}_{ip})\psi_{ij}^2\right]}{\mathbf{1}^\top\boldsymbol{\psi}_i}+\frac{1}{2}\sum_{i=1}^{K}\partial_{\delta_i}\bar{x}_{ip}\end{bmatrix}\\&=-\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)\psi_{ij}^2}{\mathbf{1}^\top\boldsymbol{\psi}_i}+\begin{bmatrix}\frac{1}{2}\sum_{i=1}^{K}\partial_{\delta_i}\bar{x}_{i1}\\\vdots\\\frac{1}{2}\sum_{i=1}^{K}\partial_{\delta_i}\bar{x}_{ip}\end{bmatrix}\end{aligned}\tag{A.6}$$

Next, write

$$\bar{\boldsymbol{x}}_i = \frac{\boldsymbol{X}_i^\top \boldsymbol{\psi}_i}{\mathbf{1}^\top \boldsymbol{\psi}_i} = \sum_{j=1}^{n_i} \frac{\boldsymbol{x}_{ij}\psi_{ij}}{\mathbf{1}^\top \boldsymbol{\psi}_i} \ .$$

Then,

$$\begin{aligned}
\partial_{\delta_i} \bar{\boldsymbol{x}}_i &= \sum_{j=1}^{n_i} \frac{(\mathbf{1}^\top \boldsymbol{\psi}_i)\boldsymbol{x}_{ij}\psi_{ij}(1+\psi_{ij}) - \boldsymbol{x}_{ij}\psi_{ij}\mathbf{1}^\top \{\boldsymbol{\psi}_i \odot (\mathbf{1}+\boldsymbol{\psi}_i)\}}{(\mathbf{1}^\top \boldsymbol{\psi}_i)^2} \\
&= \begin{bmatrix} \sum_{j=1}^{n_i} \frac{(\mathbf{1}^\top \boldsymbol{\psi}_i)x_{ij1}\psi_{ij}(1+\psi_{ij}) - x_{ij1}\psi_{ij}\mathbf{1}^\top \{\boldsymbol{\psi}_i \odot (\mathbf{1}+\boldsymbol{\psi}_i)\}}{(\mathbf{1}^\top \boldsymbol{\psi}_i)^2} \\ \vdots \\ \sum_{j=1}^{n_i} \frac{(\mathbf{1}^\top \boldsymbol{\psi}_i)x_{ijp}\psi_{ij}(1+\psi_{ij}) - x_{ijp}\psi_{ij}\mathbf{1}^\top \{\boldsymbol{\psi}_i \odot (\mathbf{1}+\boldsymbol{\psi}_i)\}}{(\mathbf{1}^\top \boldsymbol{\psi}_i)^2} \end{bmatrix} = \begin{bmatrix} \partial_{\delta_i}\bar{x}_{i1} \\ \vdots \\ \partial_{\delta_i}\bar{x}_{ip} \end{bmatrix}
\end{aligned}$$

This shows that we can rewrite (A.6) as

$$\begin{aligned}
\boldsymbol{Q} &= -\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{n_i} \frac{(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)\psi_{ij}^2}{\mathbf{1}^\top \boldsymbol{\psi}_i} + \frac{1}{2}\sum_{i=1}^{K} \partial_{\delta_i}\bar{\boldsymbol{x}}_i \\
&= -\frac{1}{2}\sum_{i=1}^{K} \frac{\boldsymbol{X}_{ci}^\top \boldsymbol{\psi}_i^2}{\mathbf{1}^\top \boldsymbol{\psi}_i} + \frac{1}{2}\sum_{i=1}^{K} \frac{(\mathbf{1}^\top \boldsymbol{\psi}_i)\boldsymbol{X}_i^\top [\boldsymbol{\psi}_i \odot (\mathbf{1}+\boldsymbol{\psi}_i)]}{(\mathbf{1}^\top \boldsymbol{\psi}_i)^2} - \frac{1}{2}\sum_{i=1}^{K} \frac{(\boldsymbol{X}_i^\top \boldsymbol{\psi}_i)\mathbf{1}^\top [\boldsymbol{\psi}_i \odot (\mathbf{1}+\boldsymbol{\psi}_i)]}{(\mathbf{1}^\top \boldsymbol{\psi}_i)^2}
\end{aligned}$$

Then the adjusted profile estimating function is given by

$$\boldsymbol{g}_{adj} = \boldsymbol{g} + \boldsymbol{Q}$$

which simplifies to (2.45). □

Bibliography

Ananth, C. V., Demissie, K., and Hanley, M. L. Birth weight discordancy and adverse perinatal outcomes among twin gestations in the United States: the effect of placental abruption. *American Journal of Obstetrics and Gynecology*, 188:954–960, 2003.

Anderson, J. A. Separate sample logistic discrimination. *Biometrika*, 59:19–35, 1972.

Bahadur, R.R. A representation of the joint distribution of responses to n dichotomous outcomes. In Solomon, H., editor, *Studies in Item Analysis and Prediction*, volume IV of *Stanford Mathematical Studies in the Social Sciences*, pages 158–68. Stanford University Press., 1961.

Barndorff-Nielsen, O. On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, 70:343–365, 1983.

Bowden, J., Thompson, J.R., and Burton, P.R. A two-stage approach to the correction of ascertainment bias in complex genetic studies involving variance components. *Annals of Human Genetics*, 71:220–229, 2007.

Breslow, N. E. Covariance analysis of censored survival data. *Biometrics*, 30(1):89–99, 1974.

Breslow, N. E. and Day, N. E. *Statistical Methods in Cancer Research. Vol I. The Analysis of Case-control Studies*. IARC Scientific Publications, 1980.

Burton, P.R., Palmer, L.J., Jacobs, K., Keen, K.J., Olson, J.M., and Elston, R.C. Ascertainment adjustment: where does it take us? *The American Journal of Human Genetics*, 67: 1505–1514, 2000.

Cai, J., Qaqish, B., and Zhou, H. Marginal analysis for cluster-based case-control studies. *Sankhyā: The Indian Journal of Statistics, Series B*, 63(3):326–337, 2001.

Carey, L.A., Perou, C.M., Livasy, C.A., Dressler, L.G., Cowan, D., Conway, K., Karaca, G., Troester, M.A., Tse, C.K., Edmiston, S., et al. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *The Journal of the American Medical Association*, 295(21):2492, 2006.

Clayton, D. Conditional likelihood inference under complex ascertainment using data augmentation. *Biometrika*, 90:976, 2003.

Cornfield, J. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, 11:1269–1275, 1951.

Cox, D.R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Cox, D.R. and Reid, N. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B*, pages 1–39, 1987.

Cummings, P., McKnight, B., and Greenland, S. Matched cohort methods for injury research. *Epidemiologic Reviews*, 25:43, 2003.

Davidov, O. and Zelen, M. Referent sampling, family history and relative risk: the role of length-biased sampling. *Biostatistics*, 2(2):173, 2001.

Davis, L.J. Generalization of the Mantel-Haenszel estimator to nonconstant odds ratios. *Biometrics*, 41(2):487–495, 1985.

Diggle, P.J., Heagerty, P., Liang, K.Y., and Zeger, S.L. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

Elston, RC and Sobel, E. Sampling considerations in the gathering and analysis of pedigree data. *American Journal of Human Genetics*, 31(1):62–69, 1979.

Epstein, M.P. Comment on Ascertainment adjustment in complex diseases. *Genetic Epidemiology*, 23(3):209–213, 2002.

Epstein, M.P., Lin, X., and Boehnke, M. Ascertainment-adjusted parameter estimates revisited. *The American Journal of Human Genetics*, 70:886–895, 2002.

Evans, L. Double pair comparison–a new method to determine how occupant characteristics affect fatality risk in traffic crashes. *Accident Analysis & Prevention*, 18:217–227, 1986a.

Evans, L. The effectiveness of safety belts in preventing fatalities. *Accident Analysis & Prevention*, 18:229–241, 1986b.

Evans, L. and Frick, M.C. Helmet effectiveness in preventing motorcycle driver and passenger fatalities. *Accident Analysis & Prevention*, 20(6):447–458, 1988.

Fleiss, J. L. Review papers: The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, 2:121, 1993.

Glidden, D.V. and Liang, K.Y. Ascertainment adjustment in complex diseases. *Genetic Epidemiology*, 23:201–208, 2002.

Godambe, V.P. Orthogonality of estimating functions and nuisance parameters. *Biometrika*, 78:143, 1991.

Goetgeluk, S. and Vansteelandt, S. Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, 64:772–780, 2008.

Greenland, S. Modeling risk ratios from matched cohort data: an estimating equation approach. *Journal of the Royal Statistical Society. Series C*, 43:223–232, 1994.

Heagerty, P.J. Marginally Specified Logistic-Normal Models for Longitudinal Binary Data. *Biometrics*, 55(3):688–698, 1999.

Heagerty, P.J. and Kurland, B.F. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88:973–985, 2001.

Heagerty, P.J. and Zeger, S.L. Marginalized multilevel models and likelihood inference. *Statistical Science*, pages 1–19, 2000.

Horvitz, D.G. and Thompson, DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Hudson, J.I., Laird, N.M., and Betensky, R.A. Multivariate logistic regression for familial aggregation of two disorders. I. Development of models and methods. *American Journal of Epidemiology*, 153(5):500–505, 2001.

Jewell, N.P. *Statistics For Epidemiology*. CRC Press, 2004.

Laird, N.M. and Cuenco, K.T. Regression methods for assessing familial aggregation of disease. *Statistics in medicine*, 22(9):1447–1455, 2003.

Lawless, J.F., Kalbfleisch, J.D., and Wild, C.J. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):413–438, 1999.

Lee, W., Shi, J.Q., and Lee, Y. Approximate conditional inference in mixed-effects models with binary data. *Computational Statistics & Data Analysis*, 2009.

Liang, K. Y. and Zeger, S. L. Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science*, 10:158–173, 1995.

Liang, K.Y. Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models. *Biometrics*, 43(2):289–299, 1987.

Liang, K.Y. and Beaty, T.H. Statistical designs for familial aggregation. *Statistical Methods in Medical Research*, 9(6):543, 2000.

Liang, K.Y. and Pulver, A.E. Analysis of case-control/family sampling design. *Genetic Epidemiology*, 13:253–270, 1996.

Liang, K.Y. and Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13, 1986.

Lindsay, B. Conditional score functions: some optimality results. *Biometrika*, 69:503, 1982.

Localio, A.R., Berlin, J.A., Ten Have, T.R., and Kimmel, S.E. Adjustments for center in multicenter studies: an overview. *Annals of internal medicine*, 135:112, 2001.

Magnus, J.R. and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics, revised edition*. John Wiley, Chichester, 1999.

Manski, C.F. and Lerman, S.R. The estimation of choice probabilities from choice based samples. *Econometrica*, 45:1977–1988, 1977.

Manski, C.F. and McFadden, D. Alternative estimators and sample designs for discrete choice analysis. *Structural analysis of discrete data with econometric applications*, pages 2–50, 1981.

McCullagh, P. and Nelder, J.A. *Generalized Linear Models*. Chapman & Hall/CRC, 1989.

Meester, S.G. and Mackay, J. A parametric model for cluster correlated categorical data. *Biometrics*, 50:954–963, 1994.

Neuhaus, J., Scott, A. J., and Wild, C. J. The analysis of retrospective family studies. *Biometrika*, 89:23–37, 2002.

Neuhaus, J.M. and Jewell, N.P. The effect of retrospective sampling on binary regression models for clustered data. *Biometrics*, 46:977–990, 1990.

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*, 68:859–872, 2006.

Neuhaus, J.M., Hauck, W.W., and Kalbfleisch, J.D. The effects of mixture distribution misspecification when fitting mixed effects logistic models. *Biometrika*, 79:755–762, 1992.

Neuhaus, J.M., Scott, A.J., and Wild, C.J. Family-specific approaches to the analysis of case-control family data. *Biometrics*, 62:488–494, 2006.

Newman, B., Moorman, P.G., Millikan, R., Qaqish, B.F., Geradts, J., Aldrich, T.E., and Liu, E.T. The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology. *Breast Cancer Research and Treatment*, 35(1):51–60, 1995.

Neyman, J. and Scott, E.L. Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32, 1948.

Noh, M., Lee, Y., and Pawitan, Y. Robust ascertainment-adjusted parameter estimation. *Genetic epidemiology*, 29:68, 2005.

O'Brien, K.M., Cole, S.R., Tse, C.K., Perou, C.M., Carey, L.A., Foulkes, W.D., Dressler, L.G., Geradts, J., and Millikan, R.C. Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. *Clinical Cancer Research*, 16(24):6100, 2010.

O'Neill, T.J. and Barry, S.C. Truncated logistic regression. *Biometrics*, 51(2):533–541, 1995.

Pepe, M.S. and Anderson, G.L. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics-Simulation and Computation*, 23(4):939–951, 1994.

Pfeiffer, R.M., Gail, M.H., and Pee, D. Inference for covariates that accounts for ascertainment and random genetic effects in family studies. *Biometrika*, 88:933, 2001.

Prentice, R. L. and Pyke, R. Logistic disease incidence models and case-control studies. *Biometrika*, 66:403–411, 1979.

Qaqish, B. F. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90(2):455–463, 2003.

Qaqish, B.F., Zhou, H., and Cai, J. On case-control sampling of clustered data. *Biometrika*, 84:983–986, 1997.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Rathouz, P.J. and Liang, K.Y. Reducing sensitivity to nuisance parameters in semiparametric models: a quasi-score method. *Biometrika*, 86:857–869, 1999.

SAS Institute Inc. *SAS/STAT® Software: Version 9.2*. SAS Institute Cary, NC, 2009.

Schildcrout, J. S. and Heagerty, P. J. On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics*, 9:735, 2008.

Scott, A. and Wild, C. Fitting logistic regression models in case-control studies with complex sampling. In Chambers, R.L. and Skinner, C.J., editors, *Analysis of Survey Data*. John Wiley and Son, 2003.

Scott, A. J. and Wild, C. J. Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, 96:3–27, 2001.

Scott, AJ and Wild, CJ. Fitting logistic regression models in stratified case-control studies. *Biometrics*, 47:497–510, 1991.

Scott, AJ and Wild, CJ. Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84(1):57, 1997.

Severini, T.A. Modified estimating functions. *Biometrika*, 89:333, 2002.

Severini, T.A. An approximation to the modified profile likelihood function. *Biometrika*, 85(2): 403, 1998.

Small, C.G. and McLeish, D.L. Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika*, 76:693–703, 1989.

Small, C.G. and McLeish, D.L. *Hilbert Space Methods In Probability And Statistical Inference*. Wiley, 1994.

Thompson, E. Sampling and ascertainment in genetic epidemiology: A tutorial review. *Department of Statistics, University of Washington, Seattle, Washington*, 1993.

Wacholder, S., McLaughlin, J.K., Silverman, D.T., and Mandel, J.S. Selection of controls in case-control studies. *American Journal of Epidemiology*, 135(9):1019, 1992a.

Wacholder, S., Silverman, D.T., McLaughlin, J.K., and Mandel, J.S. Selection of controls in case-control studies: II. Types of controls. *American Journal of Epidemiology*, 135(9):1029, 1992b.

Wand, M. P. Vector differential calculus in statistics. *American statistician*, 56:55–62, 2002.

Wang, M. and Hanfelt, J.J. Adjusted profile estimating function. *Biometrika*, 90:845, 2003.

Wang, M. and Hanfelt, J.J. Robust modified profile estimating function with application to the generalized estimating equation. *Journal of Statistical Planning and Inference*, 138: 2029–2044, 2008.

Wang, M. and Hanfelt, J.J. A robust method for finely stratified familial studies with proband-based sampling. *Biostatistics*, 10:364, 2009.

Waterman, R.P. and Lindsay, B.G. Projected score methods for approximating conditional scores. *Biometrika*, 83:1, 1996.

Whittemore, A.S. Logistic regression of family data from case-control studies. *Biometrika*, 82: 57–67, 1995.

Wooldridge, J.M. Asymptotic Properties of Weighted M-estimators for variable probability samples. *Econometrica*, 67:1385–1406, 1999.

Wooldridge, J.M. Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric Theory*, 17:451–470, 2001.

Zhao, L.P., Hsu, L., Holte, S., Chen, Y., Quiaoit, F., and Prentice, R.L. Combined association and aggregation analysis of data from case-control family studies. *Biometrika*, 85:299–315, 1998.