# Improved Generalized Estimating Equations For Incomplete Longitudinal Binary Data, Covariance Estimation In Small Samples, And Ordinal Data

by
Jamie Perin

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, School of Public Health.

Chapel Hill
2009

Approved by:

Dr. John Preisser, Advisor
Dr. Ceib Philipps, Committee Member
Dr. Bahjat Qaqish, Committee Member
Dr. Beth Reboussin, Committee Member
Dr. Pranab K. Sen, Committee Member

**ABSTRACT**

**JAMIE PERIN: Improved Generalized Estimating Equations For Incomplete Longitudinal Binary Data, Covariance Estimation In Small Samples, And Ordinal Data.**

**(Under the direction of Dr. John Preisser.)**

The focus of this research is to improve existing methods for the marginal modeling of associated categorical outcomes. Generalized estimating equations, based on quasi-likelihood, is in wide use to make inference on marginal mean parameters, especially for categorical data. In the case that response data are not all observed, generalized estimating equations give inconsistent parameter estimates when missingness depends on observed or unobserved outcomes. Inverse-probability weighted generalized estimating equations give valid results if missingness depends only on observed outcomes, and a missingness model is correctly specified. For our first topic we propose specific forms of semi-parametric efficient estimators in marginal models when dropouts for longitudinal binary data are missing at random. The efficiency of inverse-probability weighted generalized estimating equations is also explored in this setting.

The other specific topics of concern in this research are related to extensions of generalized estimating equations that allow for modeling associations between categorical outcomes. Although associations are often considered nuisances, it is not uncommon that they are scientifically relevant. It may be of interest in this case to model associations on covariates defined by characteristics of clusters or outcome pairs. Alternating logistic regressions model marginal means of correlated binary outcomes while simultaneously allowing for an association model that parameterizes the odds ratio for outcome pairs. Our second topic concerns point and variance estimation of association parameters for finite samples. Bias adjustments in estimating outcome variance have recently been introduced for small samples in generalized estimating equations. We propose an

extension of these adjustments to odds ratio parameters in alternating logistic regressions.

The remaining topic of our research concerns generalized estimating equations for ordinal data, for which alternating logistic regressions has recently been adapted. An alternate formulation of alternating logistic regressions based on orthogonalized residuals has been introduced for binary data resolving some problems in the existing procedure, including lack of invariance of the variance estimator to observation order. In our final topic we define this alternate formulation of alternating logistic regressions for correlated ordinal data, and examine its efficiency with regards to estimating within-cluster association parameters.

# Acknowledgements

I would like to extend a special thanks to my advisor, John Preisser, for his incredible generosity with time and expertise. I feel very lucky to have been in his mentorship, and it has been a pleasure to have worked with him in my years at UNC. I would also like to thank the members of my committee, Ceib Phillips, Beth Reboussin, Bahjat Qaqish, and Pranab K. Sen, for their time and also for the insightful comments which helped to improve the quality of this research.

Furthermore, I would like to thank the National Institute of Environmental Health Sciences, for the Training Grant which allowed me to focus primarily on this work, and Lawrence Kupper and Amy Herring for providing me with that opportunity. I would also like to thank Gary Koch, who has been a very positive force in my education, in my experience at the Biometric Consulting Lab and beyond.

In addition, I would like to thank Melissa Hobgood and Tania Osborne for their encouragement. Last but not least, I would like to thank my family and friends for their love and support while this was being written, with a special thanks to Anita Abraham, who was a wise friend throughout this process.

# Contents

**List of Figures** viii

**List of Tables** ix

**1 Introduction and Literature Review** 1

1.1 Introduction . . . . . 1

1.2 Literature review for marginal modeling with estimating equations of incomplete longitudinal binary data . . . . . 3

1.2.1 Bias of generalized estimating equations and approaches when data are missing at random . . . . . 3

1.2.2 Inverse-probability weighted estimators . . . . . 7

1.2.3 Inverse-probability weighted estimators and semi-parametric efficiency . . . . . 8

1.3 Literature review for alternating logistic regressions in finite samples and for ordinal data . . . . . 11

1.3.1 Alternating logistic regressions and orthogonalized residuals . . 11

1.3.2 Estimating equation procedures with improved finite sample properties . . . . . 14

1.3.3 Alternating logistic regressions for ordinal data . . . . . 17

**2 Semi-parametric Efficient Estimation for Incomplete Longitudinal Binary Data with Application to Smoking Trends** 21

2.1 Introduction . . . . . 21

2.2 Methods . . . . . 26

2.2.1 Model . . . . . 26
2.2.2 Estimator class . . . . . 26
2.2.3 Most efficient estimator . . . . . 28
2.2.4 Efficiency of $\hat{\boldsymbol{\beta}}_G$ . . . . . 30
2.2.5 Estimation in a simple case . . . . . 30
2.2.6 Estimation in the general case . . . . . 32
2.3 Simulation . . . . . 34
2.3.1 Simple case . . . . . 34
2.3.2 Extended case . . . . . 37
2.4 Application . . . . . 39
2.4.1 Estimation of 15-year smoking trends . . . . . 39
2.4.2 Sensitivity to MAR . . . . . 41
2.5 Conclusions . . . . . 44

**3 Alternating Logistic Regressions With Improved Finite Sample Properties 54**

3.1 Introduction . . . . . 54
3.2 Finite sample corrections for ALR . . . . . 56
3.2.1 Bias-corrected estimating equations . . . . . 58
3.2.2 Bias-corrected covariance estimation . . . . . 59
3.3 Simulation . . . . . 60
3.3.1 Simulation results . . . . . 62
3.4 Example . . . . . 63
3.5 Discussion . . . . . 66

**4 Alternating Logistic Regressions for Ordinal Data 74**

4.1 Introduction . . . . . 74

4.2 Alternating logistic regressions for ordinal data . . . . . 77
4.3 Orthogonalized residuals . . . . . 79
4.4 Example . . . . . 82
4.5 Simulation . . . . . 87
4.5.1 Simulation results . . . . . 88
4.6 Conclusions . . . . . 89

**5 Summary and Future Research** **92**

5.1 Summary of research . . . . . 92
5.1.1 Semi-parametric efficient estimation for incomplete longitudinal binary data . . . . . 92
5.1.2 Alternating logistic regressions with improved finite sample properties . . . . . 93
5.1.3 Orthogonalized residuals for ordinal data . . . . . 93
5.2 Future research . . . . . 94
5.2.1 Semi-parametric efficient estimation for incomplete longitudinal binary data . . . . . 94
5.2.2 Alternating logistic regressions with improved finite sample properties . . . . . 94
5.2.3 Orthogonalized residuals for ordinal data . . . . . 95

**A Appendix** **96**

Asymptotic distribution of ORTH estimators . . . . . 96
Variance in ORTH association estimating equations . . . . . 99
ORTH binary equivalence . . . . . 100

**References** **102**

# List of Figures

2.1 The observed smoking rates among CARDIA participants over fifteen years, by ethnicity and gender . . . . . 48

2.2 The estimated difference between smoking rates at time 6 and smoking rates at time 1 in the CARDIA data . . . . . 53

3.1 Coverage of nominally 95% confidence intervals for standard ALR and ALR with proposed MMEE adjustment. . . . . 68

3.2 Odds Ratio estimates of within and between time associations in the EUDL data. . . . . 69

# List of Tables


2.1 Third and fourth order moments for multivariate binary distributions with equal first and second order moments . . . . . 48

2.2 Efficiency of $\hat{\beta}_{GT}$ and $\hat{\beta}_{WT}$ with respect to $\hat{\beta}_{AT}$ under MCAR . . . . . 49

2.3 Distribution among 1000 simulation runs of maximum observation weight $\text{Max}(\hat{w}_{it})$ . . . . . 50

2.4 Efficiency of $\hat{\beta}_{WT}$ with respect to $\hat{\beta}_{AT}$ under MAR . . . . . 51

2.5 Estimates of $\beta_6$ in the CARDIA analysis . . . . . 52

2.6 Estimates of $\beta_6$ standardized by cohort and attained education using the semi-parametric efficient estimator $\hat{\beta}_{A6}$ . . . . . 52

3.1 Bias in standard ALR versus ALR with the proposed finite sample adjustment . . . . . 67

3.2 Percent relative bias in standard ALR and in ALR with proposed adjustment . . . . . 70

3.3 Coverage of nominal 95% confidence intervals in standard and adjusted ALR . . . . . 71

3.4 Marginal mean parameter estimates for self-reported last 30-day alcohol use among youth in the EUDL community trial . . . . . 72

3.5 Association parameter estimates for different dichotomous outcomes in the EUDL data . . . . . 73

4.1 Parameter estimates and estimated standard errors with ALR and ORTH in the Sensory Retraining data . . . . . 90

4.2 Bias of ALR and ORTH estimates of $\hat{\boldsymbol{\alpha}}$ and Monte Carlo standard errors 91

4.3 Coverage of ALR and ORTH confidence intervals for $\hat{\boldsymbol{\alpha}}$ and average bias of the associated standard error estimates . . . . . 91

# Introduction and Literature Review

## 1.1 Introduction

The focus of this research is to improve existing methods for the marginal modeling of associated or clustered categorical outcomes. Clustering often arises in medical research, for example in cohort surveys following individuals over time, in clinical trials where multiple outcomes are measured for each subject, or in community based trials. The correlation between outcomes needs to be accounted for in order to make valid inference on the impact of covariates on marginal mean parameter estimates, even when this association is not scientifically relevant. In this case the dependence of outcome variables on each other is a nuisance, however, this dependence is often of direct scientific interest, for instance, in determining sample size. As the distributions of clustered binary or categorical data are generally not characterized by a small number of parameters as in Gaussian data, methods to model data without fully specifying their joint distributions have been developed.

Liang and Zeger (1986) introduced generalized estimating equations, based on the quasi-likelihood of Wedderburn (1974). Liang and Zeger's method is in wide use to make inference on marginal mean parameters, especially in the case of categorical data. This method is convenient computationally, however, it has limitations for estimating mean parameters when data are incomplete. In the case that response data are not all observed, generalized estimating equations as specified by Liang and Zeger give asymptot-

ically biased parameter estimates when missingness depends on observed or unobserved outcomes. Generalized estimating equations yield unbiased estimates only in the case that data observation is not dependent on other outcomes, or is "missing completely at random" (Little, 1988). Inverse-probability weighted generalized estimating equations give valid results if missingness depends only on observed outcomes, i.e. data is "missing at random" (Rubin, 1976) This research explores the efficiency of inverse-probability weighted generalized estimating equations when dropouts in marginal models for incomplete longitudinal binary data are missing at random and proposes specific forms of semi-parametric efficient estimators in this setting. This issue is addressed in §1.2 and §2.

The other specific topics of concern in this research are related to extensions of generalized estimating equations that allow for modeling associations between binary and categorical outcomes. Although associations are often considered nuisances, it is not uncommon that they are scientifically relevant. It may be of interest in this case to model associations on covariates defined by characteristics of clusters or outcome pairs. For example, in a community trial relating to underage drinking, it may be of interest to model the association of drinking-related outcomes based on age. Accommodating this parameterization, estimating equations have been defined for associations characterized by correlations, in addition to estimating equations characterized by odds ratios. In particular, alternating logistic regressions were defined by Carey, Zeger, and Diggle (1993) to model marginal means of correlated binary outcomes while simultaneously allowing for an association model that parameterized the odds ratio for outcome pairs.

Our second topic concerns alternating logistic regressions for finite samples. Bias adjustments in estimating the variance of $\mathbf{Y}_i$ have recently been introduced for small samples in generalized estimating equations. We propose an extension of these adjustments to alternating logistic regressions when there is a small number of clusters, a

methodology and circumstance not uncommon in community trials.

The remaining topic of our research concerns generalized estimating equations for ordinal data. Alternating logistic regressions for ordinal data was introduced by Heagerty and Zeger (1996). An alternative formulation of alternating logistic regressions for binary data was defined by Zink and Qaqish (2009) that resolves the dependence of the variance estimator on the ordering of observations within clusters. In §1.3 and §4 a formulation of ALR is defined for ordinal data based on the orthogonalized residuals of Zink and Qaqish (2009).

## 1.2 Literature review for marginal modeling with estimating equations of incomplete longitudinal binary data

### 1.2.1 Bias of generalized estimating equations and approaches when data are missing at random

Let $\mathbf{Y}_i$ be a longitudinal binary outcome, so that $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$ for binary $Y_{it}$. Our research is concerned with analysis where outcome $\mathbf{Y}_i$ is not completely observed, and the relationship between the marginal mean of $Y_{it}$ and covariate vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{iT})$ is of interest. It is often the case in longitudinal data that $\mathbf{Y}_i$ is monotonically incomplete, so that for $Y_{it}$ observed then $Y_{ij}$ is also observed for $j < t$. Let $\bar{\mathbf{Y}}_{it} := (Y_{i1}, \ldots, Y_{i(t-1)})'$ represent the history of $\mathbf{Y}_i$ at $t$. In the case that $\mathbf{Y}_i$ is monotonically incomplete, then for $Y_{it}$ observed, $\bar{\mathbf{Y}}_{it}$ is also observed. Let the marginal mean of $Y_{it}$ be restricted by

$$E(Y_{it}|\mathbf{X}_i) \;=\; g_t(\mathbf{X}_i, \boldsymbol{\beta}),\; t = 1, \ldots, T\,, \tag{1.1}$$

for known functions $g_t(\cdot, \cdot)$. For known $\mathbf{X}_i$, the marginal mean of $Y_{it}$ is then fully specified by the $p \times 1$ parameter vector $\boldsymbol{\beta}$.

Quasi-Likelihood was defined by Wedderburn (1974), in which inference is made for marginal mean parameter $\boldsymbol{\beta}$ for a given relationship between the marginal mean and variance of a random variable, without specifying a joint distribution. Liang and Zeger (1986) expanded quasi-likelihood to clustered data, in which inference for parameters related to the mean of longitudinal binary data could be made without specifying a full joint likelihood, which is computationally infeasible for cluster sizes that are not small.

The generalized estimating equations methodology proposed by Liang and Zeger (1986) has recently come into wide use to estimate marginal mean parameters such as $\boldsymbol{\beta}$ in (1.1), especially for correlated binary data. This is despite the known drawback that their parameter estimates are inconsistent for data that are observed conditional on outcome data, or are not missing completely at random (MCAR) (Preisser et al., 2002). In the case that $Y_{it}$ is not MCAR,

$$E(Y_{it}|R_{it} = 1, \mathbf{X}_i) \neq E(Y_{it}|\mathbf{X}_i) \, ,$$

where $R_{it}$ is an indicator that $Y_{it}$ is observed, so that generalized estimating equations can be biased for $E(Y_{it}|\mathbf{X}_i)$. This is in contrast to maximum likelihood, which has unbiased parameter estimates for broader missing data conditions, including data that are missing at random. While a joint likelihood can be specified for longitudinal binary data, the model complexity for large cluster sizes can be prohibitive (Preisser et al., 2000). Methods based on maximum likelihood have been developed for this case, including generalized linear mixed models (Breslow and Clayton, 1993), although the parameters are not comparable to those of marginal methods (Zeger et al., 1988). Marginal methods are used when the expectation $E(Y_{it}|\mathbf{X}_i)$, the unconditional or population level mean, is of direct interest. Our research does not include maximum likelihood methods,

instead concentrating on semi-parametric methodology. Also outside the scope of this research is data that is not missing at random, or non-ignorably missing, a scenario for which both generalized estimating equations and maximum likelihood model parameter estimates are biased.

Because generalized estimating equations are widely used and are known to have inconsistent parameter estimates for data that are missing at random, various solutions have been proposed in the literature for longitudinal binary data. Lipsitz et al. (2000) introduced an extension of generalized estimating equations valid under MCAR in which correlation parameters are estimated using a multivariate normal likelihood. Lipsitz et al. showed that when correlations of binary variates were estimated (inconsistently) with a Gaussian likelihood, as opposed to the all-available-pairs method used by Liang and Zeger (1986), the bias of the resulting estimator was reduced for data that were missing at random, for clusters of size two.

Fitzmaurice et al. (2001) compared the bias of parameter estimates when data are missing at random (MAR) for several different marginal methods valid under MCAR (but not MAR), including the Gaussian estimation described above of Lipsitz et al. (2000). The bias comparison of Fitzmaurice et al. (2001) was primarily for association parameters, although bias in mean parameter estimates was also examined, finding asymptotic bias in GEE for both mean and association parameters. Included in their comparison was an extension of generalized estimating equations proposed by Lipsitz and Fitzmaurice (1996) estimating the correlation of binary variates with conditional residuals, and second-order generalized estimated equations proposed by Liang et al. (1992).

In addition to these methods, Paik (1997) proposed an a daptation for semi-parametric inference in the case of incomplete data, where unobserved $Y_{it}$ are imputed. Paik proposed that $Y_{it}$ can be sequentially imputed using the expectation of $Y_{it}$ conditional

on the dropout pattern of $\mathbf{Y}_i$ estimated by a sample average. In the case that there is little data with exactly the same history, the conditional expectation can be modeled. For data that are MCAR or MAR, Paik's method yields estimates of $\boldsymbol{\beta}$ that are asymptotically unbiased.

Also yielding consistent estimates of $\boldsymbol{\beta}$ for MAR data in marginal models is the pattern mixture model proposed by Fitzmaurice and Laird (2000). Pattern mixture models stratify incomplete data by response pattern and model data within strata. The final model is then an average across the different patterns of incompleteness, weighted by the marginal probabilities for each pattern. Fitzmaurice and Laird (2000) used pattern mixture models to make inference in generalized estimating equations for consistent estimation of $\boldsymbol{\beta}$ (under correct marginal model specification) when data are not MCAR.

Approaches in the statistical literature that adjust GEE for consistency under MAR include both the careful estimating of outcome correlation and modeling or imputing data within missingness strata. The former approach has the advantage that correlation modeling is straightforwardly accommodated in the existing methodology and also that bias was shown empirically to be reduced for some small cluster analyses (Fitzmaurice et al., 2001; Lipsitz et al., 2000). Although bias in $\hat{\boldsymbol{\beta}}$ can be reduced for MAR data when the outcome correlation is correctly specified (Liang and Zeger, 1986), their asymptotic bias for clustered data in general is in question, especially for complex clustered data or for clusters that are not small. Even in the case that these estimators were consistent for $\boldsymbol{\beta}$, they are not necessarily the estimates with the smallest variance, as GEE estimates do not have the smallest variance. For the latter approach, imputation and modeling data within missingness strata are advantaged in their accommodation of different missing data patterns. Although imputation and pattern mixture modeling yield consistent estimators of $\boldsymbol{\beta}$, they require assumptions

about conditional distributions that are not of direct interest. In multiple imputation, it is not necessarily obvious that a multivariate distribution exists that accommodates both the conditional means and the marginally specified mean and covariance (Paik, 1997). A disadvantage for pattern mixture models is that the natural parameters of interest are not directly available (Fitzmaurice and Laird, 2000). A disadvantage for both pattern mixture models and imputation is that their estimates of marginal mean parameters are not the most efficient.

### 1.2.2 Inverse-probability weighted estimators

While Fitzmaurice and Laird (2000) modeled $\mathbf{Y}_i$ conditional on missingness, another method is a selection model approach (Little, 1995) that conversely models missingness conditional on observed $\mathbf{Y}_i$. Introduced by Robins, Rotnitzky, and Zhao (1995), this methodology proposes generalized estimating equations weighted by the inverse of the conditional observation probability, and so is called inverse-probability weighted estimating equations.

The class of estimators based on inverse-probability observation weights as introduced by Robins et al. (1995) are consistent for $\boldsymbol{\beta}$ when $\mathbf{Y}_i$ is MAR, given that a model for the observation probability of $Y_{it}$ is correctly specified. A particular estimator in this class has come into wide use as a practical adaptation to the estimator introduced by Liang and Zeger (1986) in the case that data are not MCAR, receiving considerable interest among researchers in statistical methods and practice (Troxel, 1998; Hogan et al., 2004; Preisser et al., 2000; Miller et al., 2001; Yi and Cook, 2002; Ziegler et al., 2003; Jansen et al., 2006). This estimator as introduced by Robins et al. (1995) is defined in detail in §2.2.2.

Although there is a large variety of methods available for analyzing incomplete longitudinal binary data in the statistical literature, we have focused here on efficient

estimators for $\boldsymbol{\beta}$ in semi-parametric methods. Semi-parametric models are widely used for incomplete correlated binary data, whose joint distribution can be intractable for large cluster sizes, however, the most efficient semi-parametric estimator has not been implemented. For a comprehensive review of incomplete longitudinal binary data and related analysis methods see Hogan et al. (2004).

### 1.2.3 Inverse-probability weighted estimators and semi-parametric efficiency

Efficiency in semi-parametric models when $\mathbf{Y}_i$ is completely observed was examined by Chamberlain (1987), who showed that the multivariate generalization of the quasi-likelihood estimator in the class of estimators defined by Liang and Zeger (1986) asymptotically attains the semi-parametric variance bound when the variance of $\mathbf{Y}_i$ is known. The semi-parametric variance bound for estimators of $\boldsymbol{\beta}$ is the supremum of the set of variances for all parametric submodels for the distribution of $\mathbf{Y}_i$ (Tsiatis, 2006).

The estimator defined by Liang and Zeger (1986) likewise attains the semi-parametric variance bound for complete data, under mild regulatory conditions, when substituting an estimate of the variance of $\mathbf{Y}_i$, as shown by Newey (1990). This is equivalent to Liang and Zeger's estimator when the structure of the covariance matrix is correctly specified, so that their estimator is asymptotically the most efficient when $\mathbf{Y}_i$ is completely observed. An efficient estimator for complete data was also specified by Qu et al. (2000), based on a modified generalized estimating equations using a decomposition of the working covariance matrix. This estimator's efficiency was maintained even when the covariance of $\mathbf{Y}_i$ was misspecified.

For longitudinal data, however, it is not uncommon that $\mathbf{Y}_i$ is incomplete. For example, in a cohort survey over $T$ observation times, a subject may be lost to follow-up at time $t \leq T$. When data are incomplete, the process determining missingness can

be relevant to the validity or efficiency of inference on $\boldsymbol{\beta}$ under restriction (1.1). The process that determines whether $\mathbf{Y}_i$ is observed at $t$ and $\mathbf{Y}_i$ itself may be independent conditional on covariate $\mathbf{X}_i$. This condition is commonly known as "missing completely at random" or MCAR. Let $R_{it}$ be an indicator that $\mathbf{Y}_i$ is observed at $t$. Formally, for conditional observation probability is defined

$$\lambda_{it} := P(R_{it} = 1 | R_{i(t-1)} = 1, \mathbf{Y}_i, \mathbf{X}_i) \, .$$

The MCAR condition is equivalent to a restriction on $\lambda_{it}$, or that

$$\lambda_{it} = P(R_{it} = 1 | R_{i(t-1)} = 1, \mathbf{X}_i) \, , \; t = 2, \ldots, T \, .$$

It is also possible that $R_{it}$ and $Y_{it}$ are independent conditional on previously observed $\mathbf{Y}_i$, or history $\bar{\mathbf{Y}}_{it}$. This condition is commonly known as "missing at random" or MAR, identified by Rubin (1976). The MAR condition is equivalent to the restriction on $\lambda_{it}$ that

$$\lambda_{it} = P(R_{it} = 1 | R_{i(t-1)} = 1, \mathbf{X}_i, \bar{\mathbf{Y}}_{it}) \, , \; t = 2, \ldots, T \, .$$

The MCAR condition is more restrictive than MAR, so that MAR data are not necessarily MCAR. It is also possible that the conditional observation of $Y_{it}$ is dependent on its value, meaning that $\mathbf{Y}_i$ is not MCAR or MAR. In this case $Y_{it}$ is "non-ignorably missing", or "not missing at random" (NMAR), and

$$P(R_{it} = 1 | R_{i(t-1)} = 1, \mathbf{Y}_i, \mathbf{X}_i) \neq P(R_{it} = 1 | R_{i(t-1)} = 1, \mathbf{X}_i, \bar{\mathbf{Y}}_{it}) \, ,$$

for $t = 2, \ldots, T$ .

When parts of response vector $\mathbf{Y}_i$ are missing completely at random (MCAR), the

estimator defined by Qu et al. (2000) has the minimum asymptotic variance among all consistent estimators in the class of linear unbiased estimating functions. The estimator defined by Liang and Zeger (1986) also has the minimum asymptotic variance in this class as long as the covariance structure is correctly specified. However, when data are MCAR, there exists an estimator under model (1.1) that improves on GEE by exploiting information available via a missing data model (Robins and Rotnitzky, 1995).

This estimator is in the class of estimators identified by Robins et al. (1995) which are consistent for $\boldsymbol{\beta}$ under MAR. The estimator identified by Robins and Rotnitzky (1995) has an asymptotic variance that attains the semi-parametric bound for estimators of $\boldsymbol{\beta}$ when $\mathbf{Y}_i$ is MCAR or MAR. Although this estimator depends on unknown quantities, and hence is unavailable for analysis, Robins and Rotnitzky outlined an iterative procedure whose solution for $\boldsymbol{\beta}$ has a limiting distribution equivalent to that of the most efficient semi-parametric estimator of $\boldsymbol{\beta}$. This procedure requires that auxiliary models be chosen for certain conditional expectations relating to $Y_{it}$, $\bar{\mathbf{Y}}_{it}$, $R_{it}$ and $R_{ij}$ for $j \geq t$. Robins and Rotnitzky (1995) implemented this estimator for continuous data, using likelihood-based models for the auxiliary quantities needed in their estimation of $\boldsymbol{\beta}$. This semi-parametric efficient estimator has not been implemented for binary data.

## 1.3 Literature review for alternating logistic regressions in finite samples and for ordinal data

### 1.3.1 Alternating logistic regressions and orthogonalized residuals

Consider data with $K$ clusters indexed by $i = 1, \ldots, K$. Cluster $i$ has $n_i$ binary observations denoted by $Y_{ij}, j = 1, \ldots n_i$, related to a covariate vector $\mathbf{X}_{ij}$ through

$$\text{logit}(\mu_{ij}) = \mathbf{X}'_{ij}\boldsymbol{\beta}\,, \tag{1.2}$$

where $\mu_{ij} = E(Y_{ij}|X_{ij})$. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \boldsymbol{\mu}_{in_i})'$. Also let $\Gamma_i$ represent var$(\mathbf{Y}_i)$. First order generalized estimating equations were defined by Liang and Zeger (1986) for the consistent estimation of $\boldsymbol{\beta}$, by solving

$$\mathbf{U}_{\boldsymbol{\beta}} = \sum_{i=1}^{K} D_i' V_i^{-1} \{\, Y_i - \mu_i(\boldsymbol{\beta}) \,\} = \mathbf{0}\,, \tag{1.3}$$

where $D_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}'$ and $V_i = diag(\sigma_{ijj}^{1/2})\; R_i \; diag(\sigma_{ijj}^{1/2})$, for $\sigma_{ijj} = \mu_{ij}(1-\mu_{ij})$. The matrix $R_i$ is a working correlation matrix approximating corr$(\mathbf{Y}_i)$ and $\sigma_{ijj} = \mu_{ij}(1-\mu_{ij})$.

Alternating logistic regressions was introduced by Carey, Zeger, and Diggle (1993) for correlated binary data, that estimates $\boldsymbol{\beta}$ with first order generalized estimating equations (GEE) and characterizes $\Gamma_i$ by pairwise odds ratio

$$\psi_{ijk} \;=\; \frac{P(Y_{ij}=1, Y_{ik}=1)P(Y_{ij}=0, Y_{ik}=0)}{P(Y_{ij}=1, Y_{ik}=0)P(Y_{ij}=0, Y_{ik}=1)}\,.$$

Alternating logistic regressions (ALR) models correlated binary data with (1.2) and

$$\log(\psi_{ijk}) \;=\; \mathbf{Z}'_{ijk}\boldsymbol{\alpha}\,, \tag{1.4}$$

where $\mathbf{Z}_{ijk}$ is a covariate vector for the pair of outcomes $Y_{ij}$ and $Y_{ik}$. The odds ratio $\psi_{ijk}$ is modeled through the parameter $\boldsymbol{\alpha}$, which is consistently estimated in a second set of estimating equations based on the expectations of $Y_{ij}$ conditional on $Y_{ik}$, $j < k < n_i$.

Let $\boldsymbol{\zeta}_i$ be a vector with elements $\zeta_{ijk} = E(Y_{ij}|Y_{ik})$ and $\mathbf{R}_i$ be the residual vector with elements $R_{ijk} = Y_{ij} - \zeta_{ijk}$. In alternating logistic regressions, $\boldsymbol{\alpha}$ is estimated by the solution to

$$\mathbf{U}_{\boldsymbol{\alpha},ALR} \;=\; \sum_{i=1}^{K} \partial\boldsymbol{\zeta}'_i/\partial\boldsymbol{\alpha}\, \mathrm{Diag}\{\boldsymbol{\zeta}_i(1-\boldsymbol{\zeta}_i)\}\, \mathbf{R}_i \;=\; \mathbf{0}\,. \tag{1.5}$$

The resulting estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are asymptotically joint multivariate normal, given that the $\boldsymbol{\alpha}$ model accurately represents the outcome covariance. If the outcome covariance is misspecified, the resulting estimate of $\boldsymbol{\beta}$ is still consistent. A sandwich estimator is available for the variance of $\hat{\boldsymbol{\alpha}}$. The ALR estimate of $\boldsymbol{\alpha}$ is invariant to the order of observations within cluster, however, the robust estimate of $\mathrm{var}(\hat{\boldsymbol{\alpha}})$ is not. SAS version 8.2 calculated covariance estimates for ALR by permuting the observations in $\mathbf{Y}_i$ and taking an average. In addition to this drawback, because the derivative matrix is stochastic, standard estimating equation theory is not applicable to (1.5) (Zink and Qaqish, 2009).

Also modeling $\Gamma_i$ with (1.4), Zink and Qaqish (2009) defined orthogonalized residuals, which estimate $\boldsymbol{\beta}$ using first order generalized estimating equations and estimate $\boldsymbol{\alpha}$ using estimating equations based on the expectations of cross-products $Y_{ij}Y_{ik}$ condi-

tional on $Y_{ij}$ and $Y_{ik}$, $j < k < n_i$. Define $\mu_{ijk} = E[Y_{ij}Y_{ik}]$ and

$$\sigma_{ijj} := \text{var}(Y_{ij}) = \mu_{ij}(1-\mu_{ij}) \qquad \sigma_{ijk} := \text{cov}(Y_{ij}, Y_{ik}) = \mu_{ijk} - \mu_{ij}\mu_{ik}.$$

In the framework of orthogonalized residuals (Zink and Qaqish, 2009), estimates for $\boldsymbol{\alpha}$ in (1.4) are obtained from the solution to

$$\mathbf{U}_{\boldsymbol{\alpha}} \;=\; \sum_{i=1}^{K} S_i' P_i^{-1} \mathbf{T}_i \;=\; \mathbf{0}\,, \tag{1.6}$$

where the vector $\mathbf{T}_i$ has elements $T_{ijk}$ such that

$$T_{ijk} \;=\; Y_{ij}Y_{ik} - \{\, \mu_{ijk} + b_{ijk:j}(Y_{ij} - \mu_{ij}) + b_{ijk:k}(Y_{ik} - \mu_{ik}) \,\}\,,$$

for

$$\begin{aligned} d_{ijk} &= \sigma_{ijj}\sigma_{ikk} - \sigma_{ijk}^2 \\ b_{ijk:j} &= \mu_{ijk}(1-\mu_{ik})(\mu_{ik} - \mu_{ijk})/d_{ijk}, \text{ and} \\ b_{ijk:k} &= \mu_{ijk}(1-\mu_{ij})(\mu_{ij} - \mu_{ijk})/d_{ijk}\,. \end{aligned}$$

The matrix $S_i$ is defined so that $S_i = E\left[-\partial \mathbf{T}_i / \partial \boldsymbol{\alpha}'\right]$ and $P_i$ is an approximate variance of $\mathbf{T}_i$, parameterized with an exchangeable correlation. When this exchangeable correlation is assumed to be zero, the resulting $\hat{\boldsymbol{\alpha}}$ is equivalent to that estimated with alternating logistic regressions (Zink and Qaqish, 2009). Unless otherwise noted, we use orthogonalized residuals in this case that the resulting parameter estimates are equal to those in alternating logistic regressions. The orthogonalized residuals formulation of alternating logistic regressions is preferred due to its superior analytic qualities, e.g. its variance estimate is invariant to the permutation of cluster observations (Zink and

Qaqish, 2006).

Both orthogonalized residuals and alternating logistic regressions have inefficient estimates of $\boldsymbol{\alpha}$ relative to second order estimating equations (Liang et al., 1992; Zink and Qaqish, 2009), which estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ parameters simultaneously. There is a considerable advantage computationally to the separate estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as in alternating logistic regressions, due to the inversion of matrices of order $n^2$, as opposed to inverting matrices of order $n^4$ as in second order estimating equations, for clusters of size $n$ (Carey et al., 1993). Orthogonalized residuals also has this computational advantage. However, orthogonalized residuals can also gain efficiency in $\hat{\boldsymbol{\alpha}}$ compared to alternating logistics regressions by assuming a non-diagonal structure for the covariance of $\mathbf{T}_i$ in (1.6), and because the residual $\mathbf{T}_i$ has a small correlation with $\mathbf{Y}_i$ (Zink and Qaqish, 2009).

Software for the implementation of orthogonalized residuals in SAS/IML and R has been developed and is publicly available (By et al., 2008). In addition, diagnostics for the effect of individual clusters and observations on $\hat{\boldsymbol{\alpha}}$ have been developed (Preisser et al., 2008).

Although $\hat{\boldsymbol{\alpha}}$ in orthogonalized residuals is consistent for $\boldsymbol{\alpha}$, the poor performance of empirical sandwich estimators when applied with a small number of clusters (e.g., less than 40) in GEE applications where primary interest is in $\boldsymbol{\beta}$ in (1.2) (Sharples and Breslow, 1992) may also be pertinent when the main focus is estimating $\boldsymbol{\alpha}$ with alternating logistic regressions or orthogonalized residuals.

### 1.3.2 Estimating equation procedures with improved finite sample properties

Alternating logistic regressions is generally used when there is direct interest in the association between elements of the response vector $\mathbf{Y}_i$. However, there is often also

a need to estimate the variance of $\mathbf{Y}_i$ as a nuisance in estimating the variance of $\hat{\boldsymbol{\beta}}$. Adjustments in estimating the variance of $\mathbf{Y}_i$ have recently been introduced in this setting, where it is only estimated as a nuisance parameter.

Letting the variance of $\mathbf{Y}_i$ be represented by $\Gamma_i$, the true variance of the estimator for $\boldsymbol{\beta}$ that is the solution to (1.3) is

$$\left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{K} D_i' V_i^{-1} \, \Gamma_i \, V_i^{-1} D_i\right) \left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i\right)^{-1} .$$

The matrices $D_i$ and $V_i$ are estimated as part of the model in estimating $\boldsymbol{\beta}$, however, $\Gamma_i$ is typically estimated by the observed $\hat{\Gamma}_i = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)'$. Substituting $\hat{\Gamma}_i$ for $\Gamma_i$, this estimator is consistent for $\text{var}(\hat{\boldsymbol{\beta}})$ and is commonly referenced as the "sandwich" or "robust" variance estimator (Liang and Zeger, 1986). The robust variance estimator for correlated data is known to be biased in small samples, and yields inflated test sizes by underestimating the true variance of $\hat{\boldsymbol{\beta}}$ (Mancl and DeRouen, 2001).

Likewise, $\hat{\Gamma}_i$ is consistent for $\text{var}(\mathbf{Y}_i)$ while $E(\hat{\Gamma}_i) \neq \Gamma_i$. Mancl and DeRouen (2001) proposed an adjustment for the sandwich variance estimator of $\text{var}(\hat{\boldsymbol{\beta}})$ with an alternate estimate of $\Gamma_i$ based on a Taylor series expansion of $\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i$ around $\boldsymbol{\beta}$. Let $H_{ij} = D_i(\sum_{l=1}^{K} D_l' V_l^{-1} D_l)^{-1} D_j' V_j^{-1}$, where the leverage of cluster $i$ is the matrix $H_{ii}$ (Preisser and Qaqish, 1996). The adjustment of Mancl and DeRouen substitutes

$$(I_{n_i} - H_{ii})^{-1} \hat{\Gamma}_i (I_{n_i} - H_{ii})^{-1\prime}$$

as an estimate of $\Gamma_i$ in place of $\hat{\Gamma}_i$.

Kauermann and Carroll (2001) proposed an adjustment to the estimate of $\Gamma_i$ also based on a Taylor series expansion of $\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i$, that

$$(I_{n_i} - H_{ii})^{-1/2} \hat{\Gamma}_i (I_{n_i} - H_{ii})^{-1/2\prime}$$

be used in place of $\hat{\Gamma}_i$ in the sandwich estimator of var($\hat{\boldsymbol{\beta}}$). Lu et al. (2007) compared the proposed covariance estimator adjustments of Mancl and DeRouen and Kauermann and Carroll in general and specifically for correlated binary data. They concluded that the adjustment of Mancl and DeRouen may overestimate $\Gamma_i$ in some scenarios, although test sizes were often closer to nominal with Mancl and DeRouen's adjustment due to substantial variance of estimated $\Gamma_i$.

Pan and Wall (2002) and Fay and Graubard (2001) proposed degree of freedom adjustments when estimating var($\hat{\boldsymbol{\beta}}$) in small samples to correct for inflated test sizes, however, the success of adjustments to the degrees of freedom has been limited (Lu et al., 2007; Braun, 2007). Degree of freedom adjustments will not be addressed here in relation to alternating logistic regression and orthogonalized residuals.

While these adjustments to the estimation of $\Gamma_i$ in small samples have mostly been applied as a means to estimating var($\hat{\boldsymbol{\beta}}$), they are also related to methods extending generalized estimating equations in which $\Gamma_i$ is modeled more explicitly. For correlated binary data, Prentice (1988) introduced a method to model intra-cluster correlations. Alternating logistic regressions (Carey et al., 1993) is another widely used method to model the associations in correlated binary data.

Sharples and Breslow (1992) proposed an adjustment to Prentice's method for correlated binary data in small samples. The estimating equations proposed by Prentice to estimate the intra-cluster correlation employ the residual $R_{ijk} - \rho_{ijk}$ where $R_{ijk} = \hat{r}_{ij}\hat{r}_{ik}$ and $\hat{r}_{ij} = (Y_{ij} - \hat{\mu}_{ij})/(\hat{\mu}_{ij}(1-\hat{\mu}_{ij}))^{1/2}$. Sharples and Breslow proposed that $\tilde{R}_{ijk} = r_{ij}r_{ik}/\{(1-h_{ij})(1-h_{ik})\}$ be used in place of $R_{ijk}$, where $h_{ij}$ and $h_{ik}$ are the $j$ and $k$ diagonal elements of the leverage matrix $H_{ii}$ defined by Preisser and Qaqish (1996).

Preisser et al. (2008) proposed an alternate adjustment to the Prentice estimating equations for intra-cluster correlations in small samples. Their proposed finite sample

adjustment substitutes $\tilde{R}_{ijk}$ for $R_{ijk}$ in the estimating equations for intra-cluster correlations, where $\tilde{R}_i$ has elements $\tilde{R}_{ijk} = \mathbf{G}_{ij.}\hat{\mathbf{R}}_{i.k}$. The vector $\mathbf{G}_{ij.}$ corresponds to the $j^{th}$ row of $G_i = (\, I_{n_i} - H_{ii}\,)^{-1}$ for $H_{ii} = D_i\,(\sum_{l=1}^{K} D_l' V_l^{-1} D_l)^{-1}\, D_i' V_i^{-1}$ and $\hat{\mathbf{R}}_{i.k}$ is the $k^{th}$ column of an empirical covariance matrix, such that $\hat{\mathbf{R}}_{i.k} = (\hat{r}_{i1}\hat{r}_{ik},\, \ldots\, , \hat{r}_{in_i}\hat{r}_{ik})'$.

Preisser et al. (2008) also applied finite sample corrections analogous to those of Mancl and DeRouen (2001) and Kauermann and Carroll (2001) in estimating var($\hat{\boldsymbol{\beta}}$) to the variance estimates for the parameters governing intra-cluster correlations. Preisser et al. concluded that the behavior of intra-cluster correlation estimates in small samples can be improved by bias-corrected estimating equations and sandwich variance estimates.

Intra-cluster correlations are a common method of quantifying association in $\mathbf{Y}_i$. Pairwise odds ratios are likewise a commonly used quantification for outcome association, also standing to benefit from less biased estimation in small samples. Alternating logistic regressions employ pairwise odds ratios in modeling associations, and can be useful in the case that outcome association is of direct interest and when the number of clusters is small, as in a community trial, where association can be used to determine appropriate sample sizes.

### 1.3.3 Alternating logistic regressions for ordinal data

Let $O_i$ be an ordinal measurement for $i = 1, \ldots, K$. The possible realizations of $O_i$ are defined for $O_i = c$, $c \in 1, \ldots, C+1$, so that for $C = 1$, $O_i$ is binary. For a vector of covariates $\mathbf{X}_i$, the distribution of $O_i$ is typically modeled with

$$\text{logit}\{P(O_i \leq c)\} = \delta_c + \mathbf{X}_i'\beta\,,\ c = 1, \ldots, C. \tag{1.7}$$

While the relationship between $\mathbf{X}_i$ and $P(O_i \leq c)$ may depend on response level $c$, it is more common in applications to assume that $\mathbf{X}_i$, $P(O_i \leq c)$, and hence $\boldsymbol{\beta}$ are not related to $c$. This is called the proportional odds model or assumption. Our interest is in the model defined by (1.7) for correlated responses (see (4.1) of §4.2), where the probability $P(O_{ij} \leq c)$ is modeled for ordinal outcome $O_{ij}$ in cluster $i = 1, \ldots, K$, for $j = 1, \ldots, n_i$, with each $O_{ij}$ on the scale $1, \ldots, C+1$. There are a number of modeling approaches available for analyzing correlated ordinal data, including those based on marginal methods and subject specific hierarchical models.

Likelihood-based methods for correlated ordinal data have recently been introduced. A likelihood-based model for bivariate ordinal data using the Plackett distribution was proposed by Dale (1986), and Molenberghs and Lesaffre (1994) extended Dale's method to apply to multivariate ordinal data. Lesaffre and Molenberghs (1991) proposed a likelihood-based probit model for multivariate ordinal data. Glonek and McCullagh (1995) introduced an alternate class of models for multivariate categorical data using the multivariate logistic transform of McCullagh and Nelder (1989) to analyze the dependency of the joint distribution on covariates.

In contrast to maximum likelihood methods, estimating equations for marginal models do not use the full joint distributions of ordinal outcomes to estimate the model for correlated data analogous to (1.7). A certain class of marginal model has recently come into wide use for correlated data with the advent of generalized estimating equations, as defined by Liang and Zeger (1986).

The application of generalized estimating equations to ordinal or categorical data has received considerable attention in the statistical literature to date. Liang, Zeger and Qaqish (1992) defined a marginal model for categorical data using generalized estimating equations based on response vectors and vectors of response cross-products. Lipsitz, Kim and Zhao (1994) also proposed estimating equations for clustered categor-

ical data based on the estimating equations of Liang and Zeger. Lipsitz et al. (1994) outlined the iterative estimation of the covariance for a select number of structures (exchangeable, 1-dependence, banded, and unstructured) based on a method of moments approach.

Marginal methods based on generalized estimating equations for ordinal data have also been examined by Clayton (1992), particularly in comparison to maximum likelihood. Gange, Linton, Scott, et al. (1995) applied generalized estimating equations for bivariate ordinal data, and Miller, Davis, and Landis (1993) showed that under certain assumptions generalized estimating equations estimators are equal to those of weighted least squares for correlated ordinal data.

Also based on generalized estimating equations (Liang and Zeger, 1986) alternating logistic regressions was introduced by Carey, Zeger, and Diggle (1993) in the analysis of multivariate binary data. Their method was extended to multivariate categorical data by Heagerty and Zeger (1996). Let $O_{ij}$ be represented by the indicator variables $Y_{ijc} = I(O_{ij} \leq c)\,, c = 1, \ldots, C$. For the vector $\boldsymbol{\zeta}_i$ with elements $\zeta_{i(j,k)(a,b)} = E(Y_{ija}|Y_{ikb})$ and residual vector $\mathbf{R}_i$ with elements $R_{i(j,k)(a,b)} = Y_{ija} - \zeta_{i(j,k)(a,b)}$, Heagerty and Zeger (1996) defined the estimating equations

$$\mathbf{U}_{\boldsymbol{\alpha}} \;=\; \sum_{i=1}^{K} \partial\boldsymbol{\zeta}_i'/\partial\boldsymbol{\alpha}\, \mathrm{Diag}\{\boldsymbol{\zeta}_i(1-\boldsymbol{\zeta}_i)\}\, \mathbf{R}_i\,,$$

in an adaptation of alternating logistic regressions to ordinal data. Heagerty and Zeger (1996) compared the efficiency for estimating $\boldsymbol{\alpha}$ of alternating logistic regressions and different marginal methods for categorical data, including second order estimating equations (Liang et al., 1992). In this proposal we seek a new $\mathbf{U}_{\boldsymbol{\alpha}}$ expression for ordinal outcomes that resolves certain deficiencies in the Heagerty and Zeger (1996) formulation, in particular, lack of invariance of the corresponding sandwich variance estimator

to the ordering of observations within cluster.

Second order estimating equations for ordinal data solve simultaneously for mean and association parameters. This method can be computationally burdensome for large clusters, having a matrix of dimension $Cn + C^2\binom{n}{2}$ to invert, where $n$ is the cluster size. Heagerty and Zeger (1996) also considered alternating logistic regressions and first order generalized estimating equations, where mean and association parameters are estimated separately. First order generalized estimating equations are less burdensome computationally for large clusters, and have high efficiency for correlation parameters when association is not strong.

In contrast to marginal methods, which model parameters at the population level, random effects have also been used to model correlated ordinal data for subject-specific effects. A random effects model for correlated ordinal data was proposed by Ezzet and Whitehead (1991) that is fit with Gaussian quadrature. Agresti and Lang (1993) proposed a model for ordinal data with subject-specific cutpoints that is fit using conditional maximum likelihood. A model proposed by Crouchley (1995) for correlated ordinal data assumes an underlying response variable, that specifies a full joint likelihood and yields a closed form for parameter inference and estimation. These methods using random effects to model correlated ordinal data will not be considered in full, as the focus of our research is on marginal methods. For a comprehensive review of models for correlated ordinal data, see Agresti (2003).

# Semi-parametric Efficient Estimation for Incomplete Longitudinal Binary Data with Application to Smoking Trends

## 2.1 Introduction

Although smoking rates in the United States have generally been tracked using cross-sectional surveys (Wagenknecht et al., 1998), smoking status has also been modeled based on longitudinal data, specifically from the ongoing Coronary Artery Risk Development in young Adults (CARDIA) study (Preisser et al., 2000). A special problem with analyzing the CARDIA data, common with many cohort surveys, is the level of dropout of study participants, which for some groups approaches 20% by the first follow-up visit. Although widely used methods for analyzing correlated binary data are not valid unless certain dropout conditions are met, recent advances in statistical methods have introduced consistent and efficient estimators for marginal mean parameters under a variety of conditions for missing data.

The Coronary Artery Risk Development in young Adults (CARDIA) study is a

population-based multicenter cohort study collecting data related to cardiovascular health begun in 1986. Binary smoking status (yes/no) in the CARDIA study was assessed at years 0, 2, 5, 7, 10 and 15 after study initiation. Applying weights based upon the estimated probability of dropout, Preisser et al. (2000) used inverse probability weighted estimators (Robins, Rotnitzky, and Zhao, 1995) to analyze smoking data from the CARDIA study for years 1986-1993. The results of these analyses indicated that smoking rates for white men and women were significantly declining, while changes in smoking rates for black men and women were not statistically different from zero, although estimated trends were positive.

The importance of addressing dropout for the CARDIA study population is shown in an analysis of an updated and expanded data set for years 1986-2001 (through year 15), where interest is in the marginal mean smoking prevalence model $\text{logit}[\mu_{it}] = \beta_0 + \beta_t$, $t = 1, \ldots, 6$, with $\beta_1 = 0$. A possible analysis for longitudinal binary data is to use generalized estimating equations (GEE) for parameter and standard error estimation, which ignores the dropout mechanism and assumes dropouts are missing completely at random (MCAR). A GEE analysis assuming an independent correlation structure of 5,077 young adults with a baseline (Year 0) exam, by ethnicity and gender group, yields estimates (standard errors) $\beta_6$ of $-.33(.091)$, $-.39(.076)$, $-.66(.089)$, and $-.76(.089)$ for black men, black women, white men, and white women respectively. These estimates suggest significant declines in smoking for all groups and correspond roughly to the observed differences in log odds of smoking between baseline year 0 and year 15. Figure 2.1 shows the observed smoking rates among the CARDIA subjects at all six observation times.

Although widely used for correlated binary data, there is evidence that these GEE estimators are likely biased for $\boldsymbol{\beta}$. First, in their analysis of the CARDIA data up to year 7, Preisser et al. (2000) reported that baseline smokers were more likely to

drop out than baseline non-smokers, meaning that data up to year 7 are not missing completely at random. In addition, in the logistic model $\text{logit}[\, E(R_{it})\,] = \beta_0 + \beta_1 Y_{i(t-1)}$, where $R_{it}$ is an indicator that subject $i$ is observed at $t$, and $Y_{i(t-1)}$ is last observed smoking status, $\beta_1$ estimates (standard errors) are -.15 (.044), -.17 (.043), -.35 (.057), and -.30 (.055) for black men, black women, white men, and white women respectively. That all these parameter estimates are significantly different from zero strongly suggests that the CARDIA data are not missing completely at random, implying that GEE is underestimating smoking rates, since non-smokers are more likely to be observed at the next observation time. In this paper, we extend these analyses applying inverse probability weighted estimators to the fifteen year CARDIA data.

Many applications, as in the CARDIA analysis, require inference on an outcome variable $Y_{it}$ given observed covariates, for subject $i$ over observation times $t = 1, \ldots, T$. Extensive literature is available on the analysis of data $Y_{it}$ under the restriction

$$E(Y_{it}|\mathbf{X}_i) \;\; = \;\; g_t(\mathbf{X}_i, \boldsymbol{\beta}) \tag{2.1}$$

known up to the $p \times 1$ parameter vector $\boldsymbol{\beta}$, where the function $g_t(\cdot, \cdot)$ is known for $t = 1, \ldots, T$. Analysis for $\boldsymbol{\beta}$ depends on assumptions about the distribution of $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$ given the $p \times T$ covariate matrix $\mathbf{X}_i$. Semi-parametric analysis makes minimal assumptions about this distribution, and is in wide use to quantify association between $Y_{it}$ and $\mathbf{X}_i$.

Generalized estimating equations (Liang and Zeger, 1986) extended the quasi-likelihood method of Wedderburn (1974) to provide asymptotically optimal estimation of $\boldsymbol{\beta}$ when the structure of the working or assumed covariance matrix is correctly specified and all $Y_{it}$ are observed (Chamberlain, 1987). An efficient estimator in this setting was also specified by Qu et al. (2000), based on a decomposition of the working covariance matrix, whose efficiency was maintained even when the covariance of $\mathbf{Y}_i$ was

misspecified.

When elements of response vector $\mathbf{Y}_i$ are MCAR, the estimator defined by Qu et al. (2000) has the minimum asymptotic variance among all estimators in the class of linear unbiased estimating functions. The estimators defined by Liang and Zeger (1986) also have the minimum asymptotic variance in this class as long as the covariance structure is correctly specified. However, when data are MCAR, there exists an estimator under model (2.1) in an expanded class that improves on GEE by exploiting information available via a missing data model (Robins and Rotnitzky, 1995). Additionally, this estimator belongs to a class of estimators that are consistent and asymptotically normal under the milder condition that data are missing at random (MAR) in the sense of Rubin (1976). In contrast, the estimators of Liang and Zeger and Qu et al. are consistent and asymptotically normal only under the more restrictive condition that $\mathbf{Y}_i$ is MCAR.

While relaxing the MCAR condition, and without any additional distributional assumptions about $\mathbf{Y}_i$ beyond (2.1), Robins, Rotnitzky, and Zhao (1995) proposed a class of weighted generalized estimating equations, also called inverse probability weighted estimators. The estimators in this class are consistent for $\boldsymbol{\beta}$ when data are MAR and the model for the missingness mechanism is correctly specified. Robins, Rotnitzky, and Zhao proposed a relatively computationally simple estimator within this class based on observation-level weights that has received considerable interest among researchers in statistical methods and practice (Troxel, 1998; Hogan et al., 2004; Preisser et al., 2000; Preisser et al., 2002; Miller et al., 2001; Yi and Cook, 2002; Ziegler et al., 2003; Jansen et al., 2006). An alternative computationally simple weighted GEE estimator (Fitzmaurice et al., 1995; Molenberghs and Verbeke, 2005) based on cluster-level weights has been found to be less efficient (O'Hara Hines, R. J. et al., 1999; Preisser et al., 2002).

Within the class proposed by Robins, Rotnitzky, and Zhao, Robins and Rotnitzky

(1995) defined the estimating equations whose solution for $\boldsymbol{\beta}$ has a limiting distribution equivalent to that of the most efficient semi-parametric estimator of $\boldsymbol{\beta}$. They provide simulated results for a continuous outcome, showing improved efficiency relative to the computationally simpler GEE for MCAR data.

This paper is concerned with optimal estimation for binary data under minimal distributional assumptions for $\mathbf{Y}_i$ in the presence of missing data. We introduce, and provide here in detail, the form of the computationally complex semi-parametric efficient estimator of Robins and Rotnitzky (1995) for longitudinal binary data, with specific algorithms for estimator generation, and assess its efficiency with respect to the Robins et.al. (1995) inverse probability weighted estimator. Given its complexity, the details of this estimator for a simple case with small cluster size and for a general case will be presented separately.

This paper also expands the results of an analysis of seven-year smoking trends we have previously undertaken (Preisser et al., 2000), applying the semi-parametric efficient estimator to trend data out to fifteen years. In the observation of smoking trends, the baseline smokers tended to drop out at greater rates than non-smokers (i.e., suggesting $Y_{it}$ is not MCAR), a situation where standard GEE is no longer consistent. A GEE analysis would exaggerate the rate of decline in smoking, while a weighted GEE analysis up-weights observed data from smokers at later time points and provides a larger (and more likely valid) estimate of the slope.

The semi-parametric efficient estimator for when $Y_{it}$ is a binary outcome and $\boldsymbol{\beta}$ is a parameter associated with a change in proportion over time is defined in Section 2. Details of this estimator for a simple case are presented first, followed by those for a more general case. A comparison study of estimators of $\boldsymbol{\beta}$ using simulated data is presented in Section 3. In the simulation study, we consider the generation of correlated binary data having different joint distributions of $\mathbf{Y}_i$ with the same first and second

moments. An analysis of 15-year smoking trends from 1986 to 2001 in the CARDIA data in presented in Section 4, and Section 5 provides conclusions.

## 2.2 Methods

### 2.2.1 Model

Let $Y_{it}$ be a binary measurement for subject $i = 1, \ldots, K$ at fixed measurement times $t = 1, \ldots, T$. The complete outcome for subject $i$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})'$, is a vector with $T$ elements, and the explanatory covariate vector $\mathbf{X}_i$ is completely observed. This is the case, for example, when the covariates are non-stochastic functions of known quantities such as time, or are observed at baseline. Let $\mathbf{V}_i$ be an observed auxiliary covariate vector, not included in the marginal mean model of $Y_{it}$ and whose elements are not contained in $\mathbf{X}_i$. Let $R_{it} = 1$ if subject $i$ is observed at time $t$ and $R_{it} = 0$ otherwise. The marginal mean $E(Y_{it}|\mathbf{X}_i)$ is defined by the functional form (2.1), with an unknown $p \times 1$ parameter $\boldsymbol{\beta}$.

It is also assumed that the missingness pattern of $Y_{it}$ is monotonic, so that for $R_{it} = 1$ and $1 \leq j < t$, $R_{ij} = 1$, and that all subjects are observed at $t = 1$ (i.e., $P(R_{i1} = 1) = 1$ for all $i$). The conditional probability that $Y_{it}$ is observed is $\lambda_{it} \equiv P(R_{it} = 1|R_{i(t-1)} = 1, \mathbf{W}_i)$ for $\mathbf{W}_i = (\mathbf{X}_i', \mathbf{V}_i', \mathbf{Y}_i')'$. Let $\bar{\mathbf{Y}}_{it}$ represent the history of $\mathbf{Y}_i$ at time $t$, so that $\bar{\mathbf{Y}}_{it} = (Y_{i1}, \ldots, Y_{i(t-1)})'$, and define $\bar{\mathbf{W}}_{it} = (\mathbf{X}_i', \mathbf{V}_i', \bar{\mathbf{Y}}_{it}')'$. Unless stated otherwise, we assume missingness in $Y_{it}$ is MAR, i.e.,

$$\lambda_{it} = P(R_{it} = 1|R_{i(t-1)} = 1, \mathbf{W}_i) = P(R_{it} = 1|R_{i(t-1)} = 1, \bar{\mathbf{W}}_{it}) \, , t = 2, \ldots, T \, . \quad (2.2)$$

### 2.2.2 Estimator class

Let $w_{it} = \left\{P(R_{it} = 1|\bar{\mathbf{W}}_{it})\right\}^{-1}$ and $\boldsymbol{\Delta}_i = \text{Diag}\{R_{it} w_{it}\}$. For a $p \times T$ matrix $\mathbf{D}_i(\boldsymbol{\beta})$

of arbitrary functions of $\mathbf{X}_i$ and $\boldsymbol{\beta}$, $\varepsilon_{it} = Y_{it} - g_t(\mathbf{X}_i, \boldsymbol{\beta})$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})'$, define the class of estimating equations

$$\sum_{i=1}^{K} \mathbf{D}_i(\boldsymbol{\beta})\, \boldsymbol{\Delta}_i\, \boldsymbol{\varepsilon}_i(\boldsymbol{\beta}) \;=\; 0\,, \tag{2.3}$$

indexed by $\mathbf{D}_i(\boldsymbol{\beta})$. There exists a solution to (2.3) for $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\beta}}$, that is consistent for $\boldsymbol{\beta}$ such that $\sqrt{K}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically normal under mild regulatory conditions (Robins et al., 1995).

Let $\mathbf{g}_i(\boldsymbol{\beta}) = (g_1(\mathbf{X}_i, \boldsymbol{\beta}), \ldots, g_T(\mathbf{X}_i, \boldsymbol{\beta}))'$ and let $\mathbf{C}_i$ be a working covariance matrix of $\boldsymbol{\varepsilon}_i(\boldsymbol{\beta})$ given $\mathbf{X}_i$. The estimator that solves (2.3) for $\mathbf{D}_i(\boldsymbol{\beta}) = \{\partial \mathbf{g}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\}'\, \mathbf{C}_i^{-1}$, with a correctly specified model for $\lambda_{it}$, was described by Robins et al. (1995). This estimator, valid when data is MAR, is denoted by $\hat{\boldsymbol{\beta}}_W$. The solution is obtained by iteratively reweighted least squares (IRLS) with the updated step:

$$\hat{\boldsymbol{\beta}}_W^{(r+1)} \;=\; \hat{\boldsymbol{\beta}}_W^{(r)} + \Big( \sum_{i=1}^{K} \mathbf{D}_i(\boldsymbol{\beta})\, \{\partial \mathbf{g}_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\} \Big)^{-1} \sum_{i=1}^{K} \mathbf{D}_i(\boldsymbol{\beta})\, \boldsymbol{\Delta}_i\, \boldsymbol{\varepsilon}_i(\boldsymbol{\beta})\,. \tag{2.4}$$

The estimator described by Liang and Zeger (1986) is the IRLS solution to

$$\sum_{i=1}^{K} \{\partial \mathbf{g}_i^*(\boldsymbol{\beta})/\partial \boldsymbol{\beta}\}'\, \mathbf{C}_i^{*\,-1}\, \boldsymbol{\varepsilon}_i^*(\boldsymbol{\beta}) \;=\; 0\,, \tag{2.5}$$

where the vector $\boldsymbol{\varepsilon}_i^*(\boldsymbol{\beta})$ represents the observed residuals for subject $i$, $\mathbf{C}_i^*$ is the conformable submatrix of $\mathbf{C}_i$, and $\mathbf{g}_i^*(\boldsymbol{\beta})$ the subset of $\mathbf{g}_i(\boldsymbol{\beta})$ corresponding to $\boldsymbol{\varepsilon}_i^*(\boldsymbol{\beta})$. Let $\hat{\boldsymbol{\beta}}_G$ denote the solution to (2.5). The estimator $\hat{\boldsymbol{\beta}}_G$ is consistent for $\boldsymbol{\beta}$ given that $\mathbf{Y}_i$ are MCAR, i.e., that

$$P(R_{it} = 1 | R_{i(t-1)} = 1, \mathbf{W}_i) = P(R_{it} = 1 | R_{i(t-1)} = 1, \mathbf{X}_i)\,, t = 2, \ldots, T\,. \tag{2.6}$$

Condition (2.6) implies but is more restrictive than (2.2). Under their respective missing data assumptions, $\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_W$ are consistent even under misspecified working covariance matrices.

### 2.2.3 Most efficient estimator

For $\tilde{\boldsymbol{\beta}}$ varying across different specifications of $\mathbf{D}_i(\boldsymbol{\beta})$, and given a correctly specified model for $\lambda_{it}$ governed by a $q \times 1$ parameter $\boldsymbol{\alpha}$, the $\tilde{\boldsymbol{\beta}}$ with the smallest variance in the class of estimating equations given by (2.3) is defined by the optimal $\mathbf{D}_i(\boldsymbol{\beta})$, $\mathbf{D}_i^{\text{opt}}(\boldsymbol{\beta})$. As defined in Theorem 1 in Robins and Rotnitzky (1995),

$$\mathbf{D}_i^{\text{opt}}(\boldsymbol{\beta}) = \{\partial \mathbf{g}_i / \partial \boldsymbol{\beta}\}\, \mathbf{A}_i^{-1}\,, \tag{2.7}$$

where $\mathbf{A}_i = E[\,\mathbf{U}_i\,\mathbf{U}_i'\,|\mathbf{X}_i\,]$ and $\mathbf{U}_i$ is defined by

$$\mathbf{U}_i \;=\; \boldsymbol{\Delta}_i \boldsymbol{\varepsilon}_i - \sum_{t=1}^{T} (R_{it} - \lambda_{it} R_{i(t-1)}) w_{it} \mathbf{G}_{it}\,. \tag{2.8}$$

The $T \times 1$ vector $\mathbf{G}_{it}$ has $j^{th}$ element equal to $E(\varepsilon_{ij} | R_{i(t-1)} = 1, \bar{\mathbf{W}}_{it})$ for $t \le j \le T$ and 0 for $1 \le j \le t-1$. The vector $\mathbf{U}_i$ is the weighted residual ($\boldsymbol{\Delta}_i \boldsymbol{\varepsilon}_i$) minus the projection of that residual on the space defined by the missingness model (Robins and Rotnitzky, 1995).

All estimators for $\boldsymbol{\beta}$ in the class defined by (2.3) have correctly specified models for the conditional observation probabilities $\lambda_{it}$, $t = 2, \ldots, T$. However, given a nested series of correctly specified models for $\lambda_{it}$, with increasing dimension of $\boldsymbol{\alpha}$, the asymptotic variance of an estimator of $\boldsymbol{\beta}$ that is the solution to (2.3) using a particular specification of $\mathbf{D}_i(\boldsymbol{\beta})$ is known not to increase (Robins et al., 1995). Furthermore, when $\mathbf{Q}_{it} = \mathbf{D}_i^{\text{opt}}(\boldsymbol{\beta})\, w_{it}\, \mathbf{G}_{it}$ is included as an additional covariate vector in a correctly specified model for $\lambda_{it}$, there is no further efficiency gain from additional covariates in

$\lambda_{it}$ (Robins et al., 1995). For a $p \times 1$ parameter vector $\boldsymbol{\delta}$ and $\boldsymbol{\xi} = (\boldsymbol{\alpha}', \boldsymbol{\delta}')'$, the best estimator of $\boldsymbol{\beta}$ utilizes the estimate of $\lambda_{it}$ that is modeled by

$$\text{logit}\, \lambda_{it}^{\text{opt}}(\boldsymbol{\xi}) = \text{logit}\, \lambda_{it}(\boldsymbol{\alpha}) + \boldsymbol{\delta}'\, \mathbf{Q}_{it}\ , \tag{2.9}$$

where $\lambda_{it}(\boldsymbol{\alpha})$ is a correct model for $\lambda_{it}$. The estimator $\hat{\boldsymbol{\beta}}_{opt}$ that is the solution to the estimating equation employing $\mathbf{D}_i^{\text{opt}}(\boldsymbol{\beta})$ and $\lambda_{it}^{\text{opt}}$ is asymptotically the most efficient estimator in the class of estimators defined by (2.3). Additionally, the asymptotic variance of $\hat{\boldsymbol{\beta}}_{opt}$ attains the semi-parametric variance bound for regular estimators of $\boldsymbol{\beta}$ (Robins and Rotnitzky, 1995).

Because the quantities $\mathbf{A}_i$ and $\mathbf{G}_{it}$ are not known, $\hat{\boldsymbol{\beta}}_{opt}$ is not available for data analysis; however, $\mathbf{A}_i$ and $\mathbf{G}_{it}$ can be estimated from the observed data with ancillary models. Consistency of the resulting $\boldsymbol{\beta}$ estimate does not depend on correct specification of these models; however, their quality does effect the amount of efficiency gained. Let $\hat{\mathbf{A}}_i$ and $\hat{\mathbf{G}}_{it}$ be consistent estimates of $\mathbf{A}_i$ and $\mathbf{G}_{it}$, and let $\hat{\boldsymbol{\beta}}_A$ be the solution for $\boldsymbol{\beta}$ in the estimating equations employing $\hat{\mathbf{D}}_i^{\text{opt}}(\boldsymbol{\beta}) = \{\partial \mathbf{g}_i/\partial \boldsymbol{\beta}\}\hat{\mathbf{A}}_i^{-1}$ and $\hat{\lambda}_{it}^{\text{opt}}$. The estimator $\hat{\boldsymbol{\beta}}_A$ has an asymptotic distribution equal to that of $\hat{\boldsymbol{\beta}}_{opt}$ (Robins and Rotnitzky, 1995). The variance of $\sqrt{K}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta})$ can be consistently estimated by

$$K \left\{ \sum_{i=1}^{K} \hat{\mathbf{D}}_i^{\text{opt}} \boldsymbol{\Delta}(\hat{\boldsymbol{\alpha}})\, \partial \mathbf{g}_i/\partial \boldsymbol{\beta} \right\}^{-1} \sum_{i=1}^{K} \tilde{\mathbf{U}}_i \tilde{\mathbf{U}}_i' \left\{ \left( \sum_{i=1}^{K} \hat{\mathbf{D}}_i^{\text{opt}} \boldsymbol{\Delta}(\hat{\boldsymbol{\alpha}})\, \partial \mathbf{g}_i/\partial \boldsymbol{\beta} \right)' \right\}^{-1} , \tag{2.10}$$

where $\tilde{\mathbf{U}}_i$ is defined by

$$\tilde{\mathbf{U}}_i = \hat{\mathbf{D}}_i^{\text{opt}}\, \boldsymbol{\Delta}_i(\hat{\boldsymbol{\alpha}})\, \hat{\boldsymbol{\varepsilon}}_i - \left( \sum_{i=1}^{K} \hat{\mathbf{D}}_i^{\text{opt}}\, \boldsymbol{\Delta}_i(\hat{\boldsymbol{\alpha}})\, \hat{\boldsymbol{\varepsilon}}_i\, \hat{\mathbf{P}}_i' \right) \left( \sum_{i=1}^{K} \hat{\mathbf{P}}_i\, \hat{\mathbf{P}}_i' \right)^{-1} \hat{\mathbf{P}}_i\ ,$$

and $\hat{\mathbf{P}}_i = \sum_{t=1}^{T} (R_{it} - \hat{\lambda}_{it} R_{i(t-1)})\, \partial \text{logit} \lambda_{it}/\partial \boldsymbol{\alpha}$, evaluated at $\hat{\boldsymbol{\alpha}}$ (Robins et al., 1995). Due to the inherent complexity in the estimation of the $T \times T$ matrix $\mathbf{A}_i$, different methods

are recommended for estimating $\mathbf{A}_i$ depending on $T$, the number of observation times for each subject, and the variability of $\mathbf{X}_i$ across subjects. An estimator of $\mathbf{A}_i$ needed to determine $\hat{\boldsymbol{\beta}}_A$ when $T$ is small and $\mathbf{X}_i$ is constant across subjects is described below, separately from the description of another method for estimating $\mathbf{A}_i$ given larger $T$ and for any $\mathbf{X}_i$.

### 2.2.4 Efficiency of $\hat{\boldsymbol{\beta}}_G$

While $\hat{\boldsymbol{\beta}}_A$ is asymptotically the most efficient of all regular and consistent semi-parametric estimators of $\boldsymbol{\beta}$ satisfying (2.1), attaining the semi-parametric variance bound (Robins and Rotnitzky, 1995), there are conditions regarding the distribution of $\mathbf{Y}_i$ for which $\hat{\boldsymbol{\beta}}_G$ is as efficient as $\hat{\boldsymbol{\beta}}_A$ under MCAR. Specifically, the asymptotic variances of $\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_A$ are equivalent when

$$E(Y_{ij}|\bar{\mathbf{Y}}_{it} = \bar{\mathbf{y}}_{it}) = \mu_{ij} + \text{cov}(Y_{ij}, \bar{\mathbf{Y}}_{it})\, \text{var}(\bar{\mathbf{Y}}_{it})^{-1}\, (\bar{\mathbf{y}}_{it} - \bar{\boldsymbol{\mu}}_{it}) \tag{2.11}$$

for $j \geq t$ (Robins and Rotnitzky, 1995), where $\mu_{ij} = E(Y_{ij})$ and $\bar{\boldsymbol{\mu}}_{it} = (\mu_{i1}, \ldots, \mu_{i(t-1)})'$. Equation (2.11) is a conditional linear property that relates the higher order moments of a multivariate distribution to its first and second order moments. Indeed, (2.11) is a property of the multivariate normal distribution. It follows that the asymptotic variance of $\hat{\boldsymbol{\beta}}_G$ attains the semi-parametric variance bound for MCAR data that satisfy the linearity property (2.11). While (2.11) may not hold in practical settings, except perhaps in an approximate sense, we show in section 3 that its utility is as an algorithm to flexibly generate correlated binary data in simulation studies.

### 2.2.5 Estimation in a simple case

In order to estimate $\boldsymbol{\beta}$, ancillary models for estimating $\mathbf{A}_i$ and $\mathbf{G}_{it}$ need to be

specified, in addition to estimating parameters in $\lambda_{it}$. Robins and Rotnitzky (1995) used maximum likelihood and least squares to estimate $\mathbf{A}_i$ and $\mathbf{G}_{it}$. In the case that the number of observation times $T$ is small, $\leq 4$, $\mathbf{X}_i$ is constant across subjects, and there is no auxiliary covariate $\mathbf{V}_i$, we show here that $\mathbf{G}_{it}$ and $\mathbf{A}_i$ can be estimated without using complex models.

Given a preliminary estimate of $\boldsymbol{\beta}$, $Y_{ij} - g_j(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$ can be regressed on the history of $\mathbf{Y}_i$ at $t$ to estimate $\mathbf{G}_{it}$, so that $\hat{\mathbf{G}}_{it}$ has elements

$$\begin{aligned}\hat{G}_{itj} \;&=\; \hat{E}(Y_{ij} - g_j(\mathbf{X}_i, \hat{\boldsymbol{\beta}}) | Y_{i1} = y_1, \ldots, Y_{i(t-1)} = y_{t-1}, R_{i(t-1)} = 1) = \\ &\frac{\sum_i I(Y_{i1} = y_1, \ldots, Y_{i(t-1)} = y_{t-1}) R_{ij} w_{ij}\, Y_{ij}}{\sum_i I(Y_{i1} = y_1, \ldots, Y_{i(t-1)} = y_{t-1}) R_{ij} w_{ij}} - g_j(\mathbf{X}_i, \hat{\boldsymbol{\beta}})\ ,\ t \leq j \leq T\ ,\end{aligned}$$

and $\hat{G}_{itj} = 0$ for $1 \leq j < t$.

In addition to $\hat{\mathbf{G}}_{it}$, an estimate for $\lambda_{it}$ is also needed in order to estimate $\mathbf{A}_i$. All the information in $\bar{\mathbf{Y}}_{it}$ is readily incorporated as covariates in a saturated model for $\lambda_{it}$ as long as $T$ is small, with no comparative advantage of modeling $\lambda_{it}$ with (2.9). Including all previously observed values of $\mathbf{Y}_i$ and all crossproducts of $\mathbf{Y}_i$, $\lambda_{it}^{\text{opt}}(\boldsymbol{\alpha})$ has $(2^T - 2)$ nuisance parameters. For example, consider the case where $Y_{it}$ is observed at up to four time points. Let $y_{it}^\star = 2y_{it} - 1$. The saturated model for the missingness of subject $i$ is determined by the conditional probabilities of observed response at times $t = 2, 3, 4$,

$$\begin{aligned}\text{logit}(\lambda_{i2}) \;&=\; \alpha_{2,0} + \alpha_{2,1} y_{i1}^\star \\ \text{logit}(\lambda_{i3}) \;&=\; \alpha_{3,0} + \alpha_{3,1} y_{i1}^\star + \alpha_{3,2} y_{i2}^\star + \alpha_{3,3} y_{i1}^\star y_{i2}^\star \\ \text{logit}(\lambda_{i4}) \;&=\; \alpha_{4,0} + \alpha_{4,1} y_{i1}^\star + \alpha_{4,2} y_{i2}^\star + \alpha_{4,3} y_{i3}^\star + \alpha_{4,4} y_{i1}^\star y_{i2}^\star \\ &\quad + \alpha_{4,5} y_{i1}^\star y_{i3}^\star + \alpha_{4,6} y_{i2}^\star y_{i3}^\star + \alpha_{4,7} y_{i1}^\star y_{i2}^\star y_{i3}^\star\ .\end{aligned}$$

These models for $\lambda_{it}$ and $\mathbf{G}_{it}$ allow that $\mathbf{A}_i$ be estimated with $\hat{\mathbf{A}}_i$, the sample covariance

of

$$\hat{\mathbf{U}}_i = \boldsymbol{\Delta}_i(\hat{\boldsymbol{\alpha}})\,\boldsymbol{\varepsilon}_i(\hat{\boldsymbol{\beta}}) - \sum_{t=1}^{T} \left\{ R_{it} - \lambda_{it}(\hat{\boldsymbol{\alpha}})\, R_{i(t-1)} \right\} w_{it}(\hat{\boldsymbol{\alpha}})\, \hat{\mathbf{G}}_{it} \tag{2.12}$$

across subjects $i = 1, \ldots, K$, where $w_{it}(\hat{\boldsymbol{\alpha}}) = \{\prod_{j=1}^{t} \lambda_{it}(\hat{\boldsymbol{\alpha}})\}^{-1}$. The matrix $\mathbf{D}_i^{\text{opt}}(\boldsymbol{\beta})$ can then be estimated by $\hat{\mathbf{D}}_i^{\text{opt}}(\boldsymbol{\beta}) = \{\partial \mathbf{g}_i / \partial \boldsymbol{\beta}\} \hat{\mathbf{A}}_i^{-1}$. First computing $\lambda_{it}^{\text{opt}}(\hat{\boldsymbol{\alpha}})$, the iterated solution to (2.3) substituting $\hat{\mathbf{D}}_i^{\text{opt}}(\boldsymbol{\beta})$, $\hat{\boldsymbol{\beta}}_A$, has an asymptotic distribution that is equivalent to the most efficient semi-parametric estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{opt}$.

### 2.2.6 Estimation in the general case

The procedure outlined above for estimating the component $\mathbf{A}_i$ of $\hat{\boldsymbol{\beta}}_A$ is for a relatively simple case. This section will outline in detail a procedure for estimating $\mathbf{A}_i$ under more general circumstances, for larger $T$ and $\mathbf{X}_i$ varying across subjects, resulting in an estimator, $\hat{\boldsymbol{\beta}}_A$, that is an approximation of the semi-parametric estimator. Ancillary models for estimating $\mathbf{A}_i$ and $\mathbf{G}_{it}$ need to be specified for this general case, in addition to estimating $\lambda_{it}$.

In order to estimate $\mathbf{G}_{it}$, note that elements $G_{itj}$ are expectations of residuals at times $j$ given the history over $t \leq j$, quantities which are not readily available from the model for the mean of $Y_{it}$ marginal to $\mathbf{X}_i$ or from the model for $\lambda_{it}$. It can be shown that, under MAR, $G_{itj}$ can also be written (Robins and Rotnitzky, 1995)

$$G_{itj} = E(w_{i(t-1)}^{-1} w_{ij}\, \varepsilon_{ij} | \bar{\mathbf{W}}_{it}, R_{ij} = 1) \times P(R_{ij} = 1 | \bar{\mathbf{W}}_{it}, R_{i(t-1)} = 1)\,. \tag{2.13}$$

Note that factors in the second part of (2.13) are not available from the missingness model except when $j = t$. A regression will be specified for the first factor in (2.13) and an additional model will be specified for the second factor. Note also that while there is a loss of efficiency in the resulting estimator of $\boldsymbol{\beta}$ if the models for these factors

are misspecified, the estimator remains consistent.

In the first factor, assuming that the conditional expectation is known up to an $m \times 1$ parameter vector $\boldsymbol{\tau}_{tj}$, so that

$$E(w_{i(t-1)}^{-1} w_{ij}\, \varepsilon_{ij} | \bar{\mathbf{W}}_{it}, R_{ij} = 1) = \boldsymbol{\tau}_{tj}' \bar{\mathbf{W}}_{it}\ ,\ 1 < t \leq j \leq T\ ,$$

the parameter $\boldsymbol{\tau}_{tj}$ can be estimated by weighted least squares. To estimate the second factor in (2.13), it is also assumed that the conditional probability $\zeta_{itj} = P(R_{ij} = 1 | \bar{\mathbf{W}}_{it}, R_{i(t-1)} = 1)$ is known up to an $m \times 1$ parameter vector $\boldsymbol{\chi}_{jt}$, corresponding to some vector $\bar{\mathbf{W}}_{it}^{*}$ , a subset of $\bar{\mathbf{W}}_{it}$ . This factor can then be estimated from the logistic model $\text{logit}(\zeta_{itj}) = \boldsymbol{\chi}_{tj}'\, \bar{\mathbf{W}}_{it}^{*}$ . Then $\hat{\mathbf{G}}_{it}$ has elements

$$\hat{G}_{itj} = \hat{\boldsymbol{\tau}}_{tj}'\, \bar{\mathbf{W}}_{it}\ \times\ \zeta_{itj}(\hat{\boldsymbol{\chi}}_{tj})\ ,\ 1 < t \leq j \leq T\ , \tag{2.14}$$

and $\hat{G}_{itj} = 0$ for $1 \leq j < t$. Although the model for $\hat{\boldsymbol{\beta}}_A$ does not directly specify the covariance of outcome $\mathbf{Y}_i$, this relationship is determined indirectly through the conditional expectations $E(Y_{ij}|Y_{it})$ in $G_{itj}$.

In addition to $\hat{\mathbf{G}}_{it}$, estimates for $\lambda_{it}$ and $\mathbf{A}_i$ are also needed to estimate $\boldsymbol{\beta}$. Because $\mathbf{A}_i$ is symmetric, there are $T(T+1)/2$ distinct elements of $\mathbf{A}_i$, each of which can be estimated with a univariate regression model of each element of matrix $\hat{\mathbf{U}}_i \hat{\mathbf{U}}_i'$ on a vector of cluster-level covariates included in $\mathbf{X}_i$, denoted by $\mathbf{X}_i^{(b)}$. Plugging in $\hat{\mathbf{G}}_{it}$ and preliminary estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, the vector $\mathbf{U}_i$ is predicted as in (2.12). Given two elements of $\hat{\mathbf{U}}_i$, $\hat{U}_{ij}$ and $\hat{U}_{ik}$, assume that $E[U_{ij}U_{ik}|\mathbf{X}_i^{(b)}] = \boldsymbol{\theta}_{jk}'\, \mathbf{X}_i^{(b)}$. The parameter $\boldsymbol{\theta}_{jk}$ can be estimated by least squares, so that

$$\hat{\boldsymbol{\theta}}_{jk} = \left( \sum_{i=1}^{K} \mathbf{X}_i^{(b)\prime} \mathbf{X}_i^{(b)} \right)^{-1} \sum_{i=1}^{K} \mathbf{X}_i^{(b)\prime}\ \hat{U}_{ij}\, \hat{U}_{ik}\ , \tag{2.15}$$

and each element of $\mathbf{A}_i$ is estimated by $\hat{\boldsymbol{\theta}}'_{jk}\mathbf{X}_i^{(b)}$. Then $\hat{\mathbf{D}}_i^{\text{opt}}(\boldsymbol{\beta}) = \{\partial \mathbf{g}_i/\partial \boldsymbol{\beta}\}\hat{\mathbf{A}}_i^{-1}$, and the best estimate of $\lambda_{it}$, $\lambda_{it}^{\text{opt}}$, is determined by (2.9). The solution to (2.3) substituting $\hat{\mathbf{D}}_i^{\text{opt}}(\boldsymbol{\beta})$ and $\lambda_{it}^{\text{opt}}(\hat{\boldsymbol{\xi}})$, $\hat{\boldsymbol{\beta}}_A$, is obtained by iterating through a sequence of estimation for $\hat{\boldsymbol{\beta}}_A$, $\hat{\boldsymbol{\xi}}$, $\hat{\mathbf{G}}_{it}$, $\hat{\mathbf{A}}_i$, and $\hat{\mathbf{D}}_i^{\text{opt}}(\boldsymbol{\beta})$, until convergence of $\hat{\boldsymbol{\beta}}_A$. The estimator $\hat{\boldsymbol{\beta}}_A$ has an asymptotic distribution that is equivalent to the most efficient semi-parametric estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{opt}$, given correct specification of $\mathbf{A}_i$ and $\mathbf{G}_{it}$ as indicated in Section 2.3, and we expect that nearly-correct specification of $\mathbf{A}_i$ and $\mathbf{G}_{it}$ will lead to nearly semi-parametric efficient estimators. As with $\hat{\boldsymbol{\beta}}_W$, correct specification of the model for $\lambda_{it}$ is required for consistency of $\hat{\boldsymbol{\beta}}_A$.

## 2.3 Simulation

### 2.3.1 Simple case

In conjunction with the methods for when $T$ is small and $\mathbf{X}_i$ is constant across clusters, a simulation study was conducted with the aim of demonstrating the gain in efficiency of $\hat{\boldsymbol{\beta}}_A$ over $\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_W$. For $T = 4$, one thousand replicates of $K = 1000$ response vectors $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i4})'$ were generated with exchangeable correlation $\rho = \text{corr}(Y_{it}, Y_{ij})$, for $\rho = 0$, 0.2, 0.6, or 0.7. An exchangeable structure was chosen as it is used to analyze the CARDIA data in Section 4. The generating distribution had mean $\mu_{it}$ given by

$$\text{logit}[\mu_{it}] = \beta_0 + \beta_T \left(\frac{t-1}{T-1}\right),\ t = 1, \ldots, T \tag{2.16}$$

for $\boldsymbol{\beta} = (\beta_0, \beta_T)'$ fixed at $(-0.7, 0.2)$. These parameter values would represent small increases in smoking rates over time if applicable to the CARDIA data.

For these fixed first and second moments, two separate sets of $\mathbf{Y}_i$ were generated:

one with a distribution satisfying the conditional linear restriction (2.11), generated using the algorithm of Qaqish (2003), and one with a distribution in violation of (2.11). Qaqish (2003) identified the conditional linear family of distributions of correlated binary data defined by (2.11). Given the means and correlations, the higher order moments of the distribution in the conditional linear family of distributions defined by (2.11) are fixed. This distribution family is a subset of all possible correlated binary data distributions (Qaqish, 2003).

Let $\mu_{1234} = E(Y_{i1}Y_{i2}Y_{i3}Y_{i4})$ and $\mu_{jkl} = E(Y_{ij}Y_{ik}Y_{il})$. Then for $T = 4$, given third order moments $\mu_{123}$, $\mu_{124}$, $\mu_{134}$, $\mu_{234}$ and fourth order moment $\mu_{1234}$, the specification of $\mathbf{Y}_i$ is complete. For the distribution of $\mathbf{Y}_i$ in violation of (2.11), these were chosen so that violation of (2.11) would be extreme. The moments for the generating distributions of $\mathbf{Y}_i$ are shown in Table 2.1.

In addition to generating $\mathbf{Y}_i$, missingness of $Y_{it}$ was generated by

$$\text{logit}\left[P(R_{it} = 1 | R_{i(t-1)} = 1, Y_{i(t-1)} = y_{i(t-1)})\right] \;=\; \alpha_0 + \alpha_1 y^{\star}_{i(t-1)}\,, \qquad (2.17)$$

for different amounts of average dropout at each $t$: 10% ($\alpha_0 = 2.2$), 20% ($\alpha_0 = 1.4$), and 40% ($\alpha_0 = 0.4$). Note that for dropout rates 10%, 20%, and 40%, the cumulative dropout at $T = 4$ is 27%, 49%, and 78%. The missingness mechanism was also varied across the relationship between $R_{it}$ and $Y_{i(t-1)}$, to simulate MCAR ($\alpha_1 = 0$), a weak MAR relationship ($\alpha_1 = -0.2$), or a strong MAR relationship ($\alpha_1 = -0.5$).

For each replicate of sample size $K = 1000$ clusters, $\hat{\boldsymbol{\beta}}_G$, $\hat{\boldsymbol{\beta}}_W$ and $\hat{\boldsymbol{\beta}}_A$ were computed in the following analysis model for the mean of $Y_{it}$,

$$\text{logit}[\mu_{it}] = \beta_0 + \beta_t\,, t = 1, \ldots, T\,, \qquad (2.18)$$

with $\beta_1 = 0$. The estimator $\hat{\boldsymbol{\beta}}_G$ was determined using both an exchangeable and

an independent correlation structure. For $\hat{\boldsymbol{\beta}}_W$, the correct missingness model (2.17) was used; note, that for the unrestricted means model (2.18), the solution $\hat{\boldsymbol{\beta}}_W$ does not depend upon the assumed correlation model, given that it is constant across all subjects. We focus here on parameter $\beta_4$, comparing time $T$ to time 1. Relative efficiency of $\hat{\beta}_4$ is measured by a ratio of mean squared errors, e.g., $\hat{E}[\,(\hat{\beta}_{A4} - \beta_4)^2\,] \,/\, \hat{E}[\,(\hat{\beta}_{W4} - \beta_4)^2\,]$ for estimator $\hat{\beta}_{W4}$, and similarly for $\hat{\beta}_{G4}$.

Table 2.2 provides efficiency results under MCAR. The efficiency of $\hat{\boldsymbol{\beta}}_G$ under a working independence correlation structure is poor for large $\rho$, as expected (Fitzmaurice, 1995). For $\mathbf{Y}_i$ satisfying (2.11) (case (A)) with one thousand clusters, the relative efficiencies of optimal $\hat{\boldsymbol{\beta}}_G$ using an exchangeable working correlation structure are near 100, while $\hat{\boldsymbol{\beta}}_W$ loses efficiency for increasing $\rho$ and dropout rate. For data violating (2.11) (case (B)), both $\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_W$ are inefficient, with efficiency slightly higher for $\hat{\boldsymbol{\beta}}_G$, and with efficiency losses directly related to dropout rate. For non-zero correlation and dropout rate of 40%, $\hat{\boldsymbol{\beta}}_A$ is unstable due to the difficulty of accurately estimating the weights; this explains the efficiencies in excess of 100 in Table 2.2. Numerical instability can be measured by calculating the maximum weight over all observations for a data set (Robins et al., 1995). Table 2.3 shows that $\hat{\boldsymbol{\beta}}_A$ may have considerably larger maximum weights than $\hat{\boldsymbol{\beta}}_W$ and the instability increases with dropout rate and with the magnitude of $\rho$.

The efficiency of $\hat{\boldsymbol{\beta}}_W$ with respect to $\hat{\boldsymbol{\beta}}_A$ under MAR is in Table 2.4. The relative efficiency of $\hat{\boldsymbol{\beta}}_W$ for data satisfying the condition (2.11) was higher than that for data in violation of (2.11), except for $\rho > 0.6$ and strong MAR. For data satisfying the conditional linear restriction (2.11), $\hat{\boldsymbol{\beta}}_W$ has efficiency comparable to $\hat{\boldsymbol{\beta}}_A$ for $\rho \leq 0.2$; exceptions where efficiency exceeds 100 illustrate finite sample performance where again instability was an issue (see Table 2.3). For data satisfying (2.11) and $\rho \geq 0.6$, the observed mean squared error of $\hat{\boldsymbol{\beta}}_W$ is almost always more than that of $\hat{\boldsymbol{\beta}}_A$. Generally,

the efficiency loss of $\hat{\boldsymbol{\beta}}_W$ increases as $\rho$ and dropout rate increase. For data violating (2.11), $\hat{\boldsymbol{\beta}}_W$ is inefficient for all scenarios with efficiency inversely related to dropout rate, and lowest for the combination of large correlation and dropout.

The observed relative bias of $\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_W$ in the simulations of Tables 2.2 and 2.4 (not shown) was negligible, except that $\hat{\boldsymbol{\beta}}_G$ is biased under MAR, similar to that seen by Preisser et al. (2000). The bias of $\hat{\boldsymbol{\beta}}_A$ was comparable to that of $\hat{\boldsymbol{\beta}}_W$. Simulations were also conducted using data generated by the algorithm of Emrich and Piedmonte (1991). Results for these simulations were similar to those for data generated using the algorithm of Qaqish and are not shown.

### 2.3.2 Extended case

One thousand replicate data sets each consisting of $K = 1000$ clusters having $\mathbf{Y}_i$ with cluster size $T = 6$ were randomly generated, with an exchangeable correlation structure and mean $\mu_{it}$ restricted by (2.16) and $\boldsymbol{\beta} = (\beta_1, \beta_T)'$ fixed at $(-0.7, 0.2)$. All $Y_{it}$ were generated with the algorithm of Qaqish (2003), with the exchangeable correlation $\rho = \{0, 0.2, 0.6, 0.7\}$.

Missingness of $Y_{it}$ was generated by (2.17) so that $Y_{it}$ is MAR. The parameter $\alpha_1$ was varied so that generated $R_{it}$ yielded $Y_{it}$ MCAR ($\alpha_1 = 0$), mildly MAR ($\alpha_1 = -.2$) or severely MAR ($\alpha_1 = -.5$), for 10%, 20%, and 40% average dropout at each $t$. The cumulative dropout at $T = 6$ respectively for these dropout rates is 41%, 67%, and 92%. A cluster-level binary covariate $\mathbf{V}_i$ was generated based on the attained education of CARDIA subjects. Education in the CARDIA data is correlated with smoking such that those with a college degree are less likely to have reported smoking. The auxiliary covariate $\mathbf{V}_i$ given $\mathbf{Y}_i$ was generated with means 0.1, 0.5, and 0.3 respectively for $\sum_{t=1}^{T} Y_{it} = T$, 0 or otherwise (i.e. always smoking, never smoking, or mixed).

Because there were no cluster-level covariates besides an intercept in $\mathbf{X}_i$, $\mathbf{A}_i$ was

estimated as described for a simple case, and not by (2.15). The auxiliary covariate $\mathbf{V}_i$ was not used in the analysis for data simulated under MCAR, in Table 2.2, and under MAR, data were analyzed both with and without $\mathbf{V}_i$. When $\mathbf{V}_i$ was used, $\hat{\boldsymbol{\beta}}_A$ was estimated with $\mathbf{V}_i$ in the model for $\lambda_{it}$ and also in the model for $G_{itj}$, as specified by (2.13), and $\hat{\boldsymbol{\beta}}_W$ used $\mathbf{V}_i$ in the model for $\lambda_{it}$.

The estimator $\hat{\boldsymbol{\beta}}_W$, fit with an exchangeable correlation structure, and with a correctly specified model for $\lambda_{it}$, is computed for all simulated data scenarios. For MCAR $Y_{it}$, the estimator $\hat{\boldsymbol{\beta}}_G$ was also calculated, with both independent and exchangeable correlation structures. Efficiency of these estimators is compared to that of $\hat{\boldsymbol{\beta}}_A$. Results comparing the relative efficiency of $\beta_6$ estimates in mean (2.16) for $Y_{it}$ are provided in Tables 2.2 and 2.4.

The relative efficiencies of $\hat{\boldsymbol{\beta}}_W$ and $\hat{\boldsymbol{\beta}}_G$ for MCAR data are in Table 2.2. For small and moderate dropout, $\hat{\boldsymbol{\beta}}_W$, $\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_A$ have similar behavior to that seen for data satisfying condition (2.11) with cluster size $T = 4$. For data satisfying condition (2.11) and $\rho \geq 0.6$, the efficiency of $\hat{\boldsymbol{\beta}}_W$ with respect to $\hat{\boldsymbol{\beta}}_A$ is lower for $T = 6$ than for $T = 4$, showing the effect of cumulatively greater dropout for $T = 6$. Under 40% dropout, the observed increase of efficiency in $\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_W$ is due to the breakdown of asymptotic behavior in $\hat{\boldsymbol{\beta}}_A$ due to instability caused by difficulty in estimating observation weights. At 40% dropout, only 8% of subjects are observed at follow-up time 6.

The relative efficiency of $\hat{\boldsymbol{\beta}}_W$ for MAR data is in Table 2.4. The efficiency of $\hat{\boldsymbol{\beta}}_W$ with $\mathbf{V}_i$ relative to $\hat{\boldsymbol{\beta}}_A$ with $\mathbf{V}_i$ is not shown, since it was very similar to the efficiency of $\hat{\boldsymbol{\beta}}_W$ without $\mathbf{V}_i$ relative to $\hat{\boldsymbol{\beta}}_A$ without $\mathbf{V}_i$. In general the efficiency of $\hat{\boldsymbol{\beta}}_W$ declines with increasing correlation between $Y_{it}$ and $Y_{ij}$, and with increasing severity and rate of dropout. Under 40% dropout and small $\rho$, $\hat{\boldsymbol{\beta}}_A$ is outperformed by $\hat{\boldsymbol{\beta}}_W$. Overall, however, these results quantify the considerable efficiency that can be gained by the additional computation in $\hat{\boldsymbol{\beta}}_A$, an approximation of the semi-parametric efficient

estimator, relative to the estimators $\hat{\boldsymbol{\beta}}_W$ and $\hat{\boldsymbol{\beta}}_G$.

In addition to the data simulated with $T = 6$, limited simulations for $T = 10$ were also run. We considered visit-specific conditional dropout rate of 10%, $\rho = \{0.2, 0.6\}$, for both MAR-weak and MAR-strong. Some instability was observed in $\hat{\boldsymbol{\beta}}_A$ for this dropout rate with $K = 1000$ clusters. However, the simulated efficiency of $\hat{\boldsymbol{\beta}}_W$ compared to $\hat{\boldsymbol{\beta}}_A$ for $T = 10$ showed a clear advantage of $\hat{\boldsymbol{\beta}}_A$ over $\hat{\boldsymbol{\beta}}_W$ at $\rho = 0.6$, in the range of 84-90% (not shown), with $K = 2000$ clusters. For this cluster size and dropout rate, approximately a third of subjects have complete data.

## 2.4 Application

### 2.4.1 Estimation of 15-year smoking trends

The adaptive estimator described in the previous section is used to analyze binary smoking status in the CARDIA study for follow-up years between 1986 and 2001, with cluster sizes as large as 6, by ethnicity/gender group. Only the monotonically missing data will be included for subjects with smoking status observed at baseline, i.e., if subject $i$ is not observed at time $t$, then observations for subject $i$ at times greater than $t$ will be ignored. A total of 1,946 person-exams out of 25,709 were omitted to create a monotone dataset. The strong assumptions needed to gain efficiency by including these data justify their omission (Robins et al., 1995).

This analysis aims to estimate the change in smoking rates within gender and ethnicity over fifteen years. For an indicator that subject $i$ was a smoker at time $t$ ($Y_{it}$), assume the mean $E(Y_{it}) = \mu_{it}$ is given by (2.18) for $t = 1, \ldots, 6$. At each $t > 1$, $\beta_t$ represents the log odds ratio of the smoking rate within group at time $t$ compared to time 1.

In order to assess the change in smoking rates between times 1 and 6, the CARDIA

data were analyzed for $t = 1, \ldots, 6$ with $\hat{\boldsymbol{\beta}}_A$ using observation weights determined by (2.9), including smoking status at the previous observation time as a predictor. Last observed smoking status had estimated coefficients (standard errors) of -0.15(.04), -0.17(.04), -0.35(.06), and -0.30(.06) in the models for $\lambda_{it}$ respectively for black men, black women, and white men and women, generally supporting MAR versus MCAR. The resulting $\hat{\beta}_{A6}$ and $\hat{\beta}_{W6}$ are presented in Table 2.5, along with analogous $\hat{\beta}_{G6}$ that assumed an exchangeable covariance structure. Correlations decay slightly over time in the CARDIA data, but not as fast as those in an autoregressive structure. An exchangeable correlation structure is used here as an approximation for the large correlations maintained over time in the CARDIA data (Preisser et al., 2000). The observation weights for $\hat{\boldsymbol{\beta}}_W$ were determined using only the previously observed outcome. Additional analysis results are provided including subject age and education ($\mathbf{V}_i$) as predictors in the missingness model for $\lambda_{it}$, and, for $\hat{\boldsymbol{\beta}}_A$, also in the model for the conditional expected value of $Y_{it}$ as described by (2.13).

As shown in Table 2.5, the estimates of $\beta_6$ for black women, and white men and women are significantly less than zero for all $\hat{\beta}_{G6}$, $\hat{\beta}_{W6}$ and $\hat{\beta}_{A6}$, indicating a decrease in smoking rates. For black men, $\hat{\beta}_{W6}$ and $\hat{\beta}_{A6}$ without $\mathbf{V}_i$ estimate a significantly negative trend, although the trend estimated by $\hat{\beta}_{W6}$ and $\hat{\beta}_{A6}$ when $\mathbf{V}_i$ is included is smaller and not significantly different from zero. When $\mathbf{V}_i$ is included, $\hat{\beta}_{A6}$ had $\max(\hat{w}_{it})$ of 3.7, 3.2, 4.0, and 5.3 for black men, black women, white men, and white women. The $\hat{\beta}_{A6}$ including auxiliary covariate $\mathbf{V}_i$ correspond to estimated decreases in the smoking rates from 1986 ($t = 1$) to 2001 ($t = 6$) of 2.4, 4.0, 6.7 and 8.7 percentage points for black men, black women, white men, and white women, respectively.

The Center for Disease Control (CDC) has also reported declines in the smoking rates for the ethnicity and gender groups in the CARDIA study for the same time period (MMWR, 1994; MMWR, 2003), although their declines are of greater magnitude. The

CDC has used cross-sectional surveys to estimate that between 1987 and 2001, the rate of smoking decreased from 28.8 to 22.8 percent in the general U.S. population. The CDC also reported a decline in the smoking rate of 10.6 percentage points for blacks, 5.1 percentage points for whites, and declines of 6.0 percent and 5.8 percent for men and women, respectively (MMWR, 1994; MMWR, 2003).

In addition to the above analysis results for the CARDIA data, estimates of $\beta_6$ were adjusted to reflect changes in smoking rates for corresponding ethnicity/gender groups of the same age nationwide. Table 2.6 has estimates of $\beta_6$ from sixteen different models, i.e., for each ethnicity/gender group, and also by age and education, each divided into two categories. These are estimated with $\hat{\boldsymbol{\beta}}_A$ including the auxiliary covariate $\mathbf{V}_i$, and the age and education standardized estimate of $\beta_6$ is determined by a weighted average of $\hat{\boldsymbol{\beta}}_A$ across age and education groups. The weights for each group were determined by the proportion of each race/gender group in the U.S. population having a college degree for each birth cohort as reported by the U.S. Census (Preisser et al., 2000). The adjusted estimates of $\hat{\beta}_{A6}$ for the U.S. population correspond to estimated decreases in the smoking rates of 2.2, 3.7, 6.7 and 9.3 percentage points for black men, black women, white men, and white women, respectively.

The variance estimator for $\hat{\boldsymbol{\beta}}_A$ in (2.10) is consistent for the true variance of $\hat{\boldsymbol{\beta}}_A$; however, the estimated variances shown in Table 2.5 are somewhat at odds with corresponding efficiencies estimated in the simulation study. Use of $\hat{\beta}_{A6}$ gave estimated variances that were between 25% and 50% smaller than estimated variances of $\hat{\beta}_{W6}$. These differences are larger than the gains of between 6% and 19% efficiency observed in the simulation scenario most similar to the CARDIA data analysis.

### 2.4.2 Sensitivity to MAR

In addition to the above analysis of the CARDIA subjects over six observation

times, the sensitivity of this analysis to the assumptions made about subject dropout is of interest. Sensitivity of $\hat{\boldsymbol{\beta}}_A$ in particular to the missing at random assumption can be gauged by assuming dropout that is not missing at random, and then estimating $\beta$ with $\hat{\boldsymbol{\beta}}_A$. As in Rotnitzky et al. (1998), the procedure used here to gauge sensitivity of $\hat{\boldsymbol{\beta}}_A$ to this assumption is based on fixing a parameter in the missingness model that corresponds to a violation of MAR and estimating the remaining parameters, and then evaluating estimates of smoking trends over a range of values for the MAR violation parameter. .

As defined in (2.2), data $\mathbf{Y}_i$ is MAR if the conditional observation probability $\lambda_{it}$ is not dependent on $Y_{it}$. If $\lambda_{it}$ depends on $Y_{it}$, which may not be observed, then the data are not MAR and

$$P(R_{it} = 1 | R_{i(t-1)} = 1, \mathbf{W}_i) \quad \neq \quad P(R_{it} = 1 | R_{i(t-1)} = 1, \bar{\mathbf{W}}_{it}) \, , \; t = 2, \ldots, 6 \, . \tag{2.19}$$

Assuming that $\lambda_{it}$ is conditionally dependent on $Y_{it}$, and also that $\lambda_{it}$ depends on history $\bar{\mathbf{Y}}_{it}$ only through the previously observed response $Y_{i(t-1)}$, then $\lambda_{it}$ can be modeled by

$$\text{logit}\{\lambda_{it}\} = \alpha_{0t} + \alpha_1 Y_{it} + \alpha_2 Y_{i(t-1)} + \boldsymbol{\alpha}_3' \mathbf{X}_i \, . \tag{2.20}$$

Hence the nature of the relationship between $\lambda_{it}$ and $Y_{it}$, and the degree to which MAR is violated, is determined by $\alpha_1$. When $\alpha_1 \neq 0$, data $Y_{it}$ are not MAR. The probabilities $\lambda_{it}$ for $t = 2, \ldots, 6$ will be estimated from the CARDIA data for a range of fixed $\alpha_1$, and used to estimate $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}_A$. This estimate $\hat{\boldsymbol{\beta}}_A$ is consistent given $\alpha_1$, and so the difference in $\hat{\boldsymbol{\beta}}_A$ from the original estimate in the previous section reflects the possible impact of data that violate the missing at random assumption.

For fixed $\alpha_1$, the regression coefficients $\alpha_{0t}$, $\alpha_2$, and $\boldsymbol{\alpha}_3$ in (2.20) cannot be estimated directly. These parameters can instead be estimated with the observed probability $\pi_{it}$,

distinct from $\lambda_{it}$, where

$$\begin{aligned}\pi_{it} &= P(R_{it}=1|R_{i(t-1)}=1, Y_{i(t-1)}, \mathbf{X}_i) \\ &= \sum_{y\in(0,1)} P(Y_{it}=y\,|R_{i(t-1)}=1, Y_{i(t-1)}, \mathbf{X}_i)\, P(R_{it}=1|R_{i(t-1)}=1, Y_{it}=y, Y_{i(t-1)}, \mathbf{X}_i)\} \,.\end{aligned}$$

For $\mu_{it}(Y_{i(t-1)}) = E(Y_{it}|R_{i(t-1)}=1, Y_{i(t-1)}, \mathbf{X}_i)$ and $\eta_t = \alpha_{0t} + \alpha_2 Y_{i(t-1)} + \boldsymbol{\alpha}_3 \mathbf{X}_i$ ,

$$\begin{aligned}\pi_{it} &= \{1-\mu_{it}(Y_{i(t-1)})\}\,\text{logit}^{-1}(\eta_t) + \mu_{it}(Y_{i(t-1)})\,\text{logit}^{-1}(\eta_t+\alpha_1) \\ &= \text{logit}^{-1}(\eta_t) + \mu_{it}(Y_{i(t-1)})\,\{\text{logit}^{-1}(\eta_t+\alpha_1) - \text{logit}^{-1}(\eta_t)\} \\ &= \text{logit}^{-1}(\eta_t)\left\{1+\mu_{it}(Y_{i(t-1)})\,b(\eta_t,\alpha_1)\right\} \,,\end{aligned}$$

where $b(\eta_t, \alpha_1) = (\exp(\alpha_1)-1)/(1+\exp(\eta_t+\alpha_1))$. So,

$$\text{logit}(\pi_{it}) = \eta_t + \log\left\{\frac{1+\mu_{it}(Y_{i(t-1)})\,b(\eta_t,\alpha_1)}{1+\exp(\eta_t)\mu_{it}(Y_{i(t-1)})\,b(\eta_t,\alpha_1)}\right\} \,. \tag{2.21}$$

Thus a logistic model with an offset holds for $\pi_{it}$, through which estimates for $\alpha_{0t}$, $\alpha_2$, and $\boldsymbol{\alpha}_3$ can be obtained iteratively. Observation weights are needed to estimate the conditional expected value $\mu_{it}(Y_{i(t-1)}) = E(Y_{it}|R_{i(t-1)}=1, Y_{i(t-1)}, \mathbf{X}_i)$, because

$$E(Y_{it}|R_{i(t-1)}=1, Y_{i(t-1)}, \mathbf{X}_i) \neq E(Y_{it}|R_{it}=1, Y_{i(t-1)}, \mathbf{X}_i)$$

by the assumption made in (2.20). Explicitly, for fixed $\alpha_1$, $\lambda_{it}$ as specified in (2.20) can be estimated with $\hat{\lambda}_{it} = \lambda_{it}(\hat{\eta}_t, \alpha_1)$ in the algorithm: (1.) Using ordinary logistic regression, estimate $\mu_{it}(Y_{i(t-1)})$ by regressing observed $Y_{it}$ on $Y_{i(t-1)}$ and $\mathbf{X}_i$. (2.) Initialize $\hat{\eta}_t$ by estimating $\pi_{it}$ without an offset. (3.) Determine value of offset in (2.21) and re-estimate $\eta_t$ and $\pi_{it}$. (4.) Estimate $\mu_{it}(Y_{i(t-1)})$ with a weighted logistic regression, using observation weights $w_{it} = 1/\hat{\lambda}_{it}$. (5.) Repeat steps 3 and 4 until convergence of $\hat{\alpha}_{0t}$, $\hat{\alpha}_2$, and $\hat{\boldsymbol{\alpha}}_3$.

Using $\hat{\alpha}_{0t}$, $\hat{\alpha}_2$, and $\hat{\boldsymbol{\alpha}}_3$ to estimate $\lambda_{it}$, $\boldsymbol{\beta}$ can be estimated with the semi-parametric efficient estimator described above in Section 2.3. In this manner, the sensitivity to the MAR assumption of the CARDIA analysis with $\hat{\boldsymbol{\beta}}_A$, incorporating auxiliary covariate $\mathbf{V}_i$, was assessed. This estimate modeled $\lambda_{it}$ with the covariates of the richest missingness model used in the analysis of the CARDIA data for $T = 6$, with smoking rate determined by (2.18) over six observation times and with observation probabilities determined by fixed $\alpha_1$. Values of $\alpha_1$ were considered from -0.5 (odds ratio $\exp(\alpha_1) = 0.61$) to 0.5 (odds ratio 1.65) by 0.1 increments, although the range of most interest is for $\alpha_1 < 0$. In the CARDIA analysis, negative $\alpha_1$ represents higher dropout for those who are smoking at time $t$, after adjusting for $\bar{\mathbf{Y}}_{it}$.

Figure 2.2 shows the difference between smoking rates at time 6 and time 1, $g_6(\mathbf{X}_i, \hat{\boldsymbol{\beta}}_A) - g_1(\mathbf{X}_i, \hat{\boldsymbol{\beta}}_A)$, with 95% confidence intervals for fixed $\alpha_1$. For all ethnicity/gender groups, the results suggest that changes in smoking rates are biased away from zero in the MAR analysis of Section 4.1, relative to the case where dropout is in fact not MAR and $\alpha_1 < 0$, i.e., where current smokers are more likely to drop out than are current non-smokers. The estimate for black men is the most sensitive to the missing at random assumption among race/gender groups, and has the widest range for $g_6(\mathbf{X}_i, \hat{\boldsymbol{\beta}}_A) - g_1(\mathbf{X}_i, \hat{\boldsymbol{\beta}}_A)$ depending on $\alpha_1$. For $\alpha_1 \leq -0.2$ (odds ratio 0.82), $\hat{\boldsymbol{\beta}}_A$ estimated a significant increase in smoking rates for black men. For black women, $g_6(\mathbf{X}_i, \hat{\boldsymbol{\beta}}_A) - g_1(\mathbf{X}_i, \hat{\boldsymbol{\beta}}_A)$ was significantly greater than zero for $\alpha_1 \leq -0.3$ (odds ratio 0.74). The analogous thresholds for white men and white women are -0.4 (odds ratio 0.67) and -0.5 (odds ratio 0.61).

## 2.5 Conclusions

The primary purpose of this paper was to define a specific form for the asymptotically semi-parametric efficient estimator, applying it to the CARDIA survey data, and to assess the efficiency of $\hat{\boldsymbol{\beta}}_W$, a computationally simple inverse probability weighted esti-

mator, relative to that estimator for longitudinal binary data with dropout. We show that there is efficiency to be gained upon $\hat{\boldsymbol{\beta}}_G$ and $\hat{\boldsymbol{\beta}}_W$ in the presence of incomplete data, although the computation of the semi-parametric efficient estimator $\hat{\boldsymbol{\beta}}_A$ is not straightforward, especially for large clusters. The percent efficiency gain depends on the nature of the data being analyzed. For small clusters, in the case where dropout rate and correlation is high, efficiency can be markedly increased even under circumstances where $\hat{\boldsymbol{\beta}}_G$ is consistent for $\boldsymbol{\beta}$.

The asymptotic distribution of $\hat{\boldsymbol{\beta}}_A$ is equivalent to that of the semi-parametric efficient estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{opt}$, but the amount of data needed to obtain a stable estimate of $\hat{\boldsymbol{\beta}}_A$ appears to be substantial. For a moderate number of small clusters ($K = 500$, not reported) some mild instability was observed for $\hat{\boldsymbol{\beta}}_A$ at 20% dropout, with additional instability observed at 40% dropout. Extrapolation suggests that $\hat{\boldsymbol{\beta}}_A$ could be practical for smaller numbers of clusters given less severe dropout (e.g., $\leq$ 10% per visit), but not for high dropout. In contrast, the simulations of Preisser et al. (2002), which were similar to the scenarios considered here, suggest that $\hat{\boldsymbol{\beta}}_W$ can be reliably used for $K =$ 200 even under severe dropout. Under MCAR, simulation results for K=1000 (Table 2) and for K=200 (Preisser et al. (2002, Table 9) suggest that $\hat{\boldsymbol{\beta}}_W$ is less efficient than GEE under a correctly specified correlation structure, at least for the models considered. The efficiency of $\hat{\boldsymbol{\beta}}_W$ may be improved under MCAR by considering auxiliary covariates in the missingness model (Robins & Rotnitzky, 1995). Whether efficiency improvements relative to $\hat{\boldsymbol{\beta}}_G$ are worth the effort under MCAR is uncertain.

A distinct disadvantage of the weighted GEE approaches considered in this paper is that they require monotone patterns of dropout, and thus do not use all the available data. One possible modification to these procedures would be to multiply impute response data at the intermittent missing time points, prior to application of the $\hat{\boldsymbol{\beta}}_W$ and $\hat{\boldsymbol{\beta}}_A$ estimators. Alternatively, under a framework of allowing for the observation

of longitudinal responses in continuous time, Lin et al. (2004) propose a class of inverse intensity-of-visit process-weighted estimators in marginal regression models that allow for arbitrary patterns of missing data. Maximum likelihood provides yet another alternative estimation approach for marginal regression models valid under MAR dropouts (Galecki et al., 2001). While this approach easily adapts to situations with intermittently missing data and does not require estimating the missing data model, it is not computationally feasible for larger cluster sizes without strong assumptions on higher order moments. Finally, maximum likelihood random effects models for binary data are popular (Breslow and Clayton, 1993), although they have parameters with subject-specific interpretations as opposed to the marginal interpretations of the models considered here. For a general review of handling dropout in longitudinal studies, see Hogan et al. (2004).

A second contribution of this paper was to extend a previously undertaken analysis of smoking trends among young adults from 7 years to 15 years. The new analysis of 15 year smoking trends confirmed the violation of the MCAR assumption and the accompanying importance of accounting for dropout in a frequentist analysis. In contradiction to CDC reports, this new analysis found that the decrease in smoking from 1987 to 2001 was more extreme for whites than blacks. The 15 year decline in smoking among the black cohort was estimated to be about 3 percentage points in this article, whereas the CDC reports a cross-sectional decline greater than 10 percentage points over the same period. Furthermore, examination of the assumption that dropout is MAR suggests that the actual decline in smoking rates in the United States may have been smaller than estimated by the new analysis reported here, drawing a less optimistic conclusion from a public health perspective than that offered by the CDC.

For $T = 6$, simulation results under MAR and (2.11) are notably discordant with the CARDIA analysis, where the estimated gain in precision from $\hat{\boldsymbol{\beta}}_A$ relative to $\hat{\boldsymbol{\beta}}_W$

exceeded expectations based upon simulated efficiency gain. This discrepancy may be due to the nature of the CARDIA data, which does not necessarily satisfy (2.11). An analysis of the first four observation times in the CARDIA study (not shown) yielded an estimated efficiency gain similar to that observed in the simulations for $T = 4$.

In sum, given evidence for use of an inverse probability weighted estimator relative to standard GEE, the issue becomes choosing an estimator in the class described by (2.3). For correlated binary data consisting of a moderate number of clusters, i.e. in the hundreds, the relatively simple, inefficient inverse probability weighted estimator of Robins, Rotnitzky and Zhao (1995) will continue to be an attractive and feasible alternative to GEE. For data consisting of a large amount of clusters numbering in the thousands, the more computationally complex semi-parametric efficient estimator of Robins and Rotnitzky (1995) may be worth the effort in terms of efficiency gain in scenarios with high dropout and large intra-cluster correlation.

In the case that the semi-parametric efficient estimator is of interest, auxiliary models need to be specified. Although consistency of $\hat{\boldsymbol{\beta}}_A$ is not jeopardized by mis-specification of these models, the associated efficiency gain of $\hat{\boldsymbol{\beta}}_A$ may be diminished. Correctly specified models maximize the amount of efficiency gained relative to more accessible estimators.

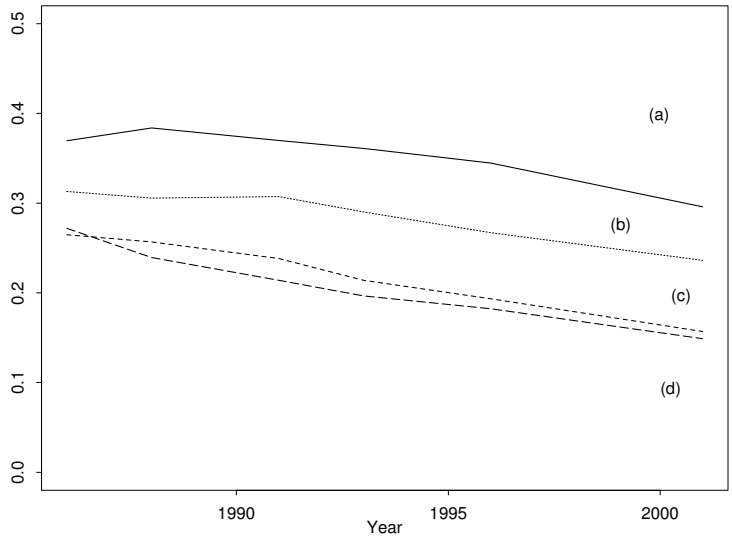


Figure 2.1: The observed smoking rates among CARDIA participants over fifteen years, by ethnicity and gender: (a) is for black men, (b) is for black women, (c) is for white men, and (d) is for white women. These data are based on 23,763 exams from 5,077 young adults (4.68 exams/person). A total of 1,946 observed exams that occurred after a missed exam were omitted to create a monotone missingness pattern needed to accommodate the weighted GEE methodology. Including all observed exams, results in GEE-independence estimates of the log odds ratio of smoking in 2001 versus 1986 of -0.19 (0.075), -0.29 (0.063), -0.57(0.079), and -0.72 (0.079) for black men, black women, white men and white women respectively.

Table 2.1: Third and fourth order moments for the multivariate binary distributions used to generate $Y_i$ in the simulation experiment for $T = 4$. Distribution (A) satisfies condition (2.11), while distribution (B) is in violation of (2.11).

| | $\rho = 0$ | | $\rho = 0.2$ | | $\rho = 0.6$ | | $\rho = 0.7$ | |
|---|---|---|---|---|---|---|---|---|
| | (A) | (B) | (A) | (B) | (A) | (B) | (A) | (B) |
| $\mu_{123}$ | 0.04 | 0.10 | 0.09 | 0.06 | 0.22 | 0.18 | 0.25 | 0.22 |
| $\mu_{124}$ | 0.04 | 0.10 | 0.10 | 0.06 | 0.22 | 0.18 | 0.25 | 0.23 |
| $\mu_{134}$ | 0.05 | 0.11 | 0.10 | 0.06 | 0.23 | 0.19 | 0.26 | 0.23 |
| $\mu_{234}$ | 0.05 | 0.11 | 0.10 | 0.06 | 0.23 | 0.19 | 0.26 | 0.23 |
| $\mu_{1234}$ | 0.02 | 0.10 | 0.06 | 0.00 | 0.20 | 0.11 | 0.24 | 0.18 |

Table 2.2: Efficiency of $\hat{\beta}_{GT}$ and $\hat{\beta}_{WT}$ with respect to $\hat{\beta}_{AT}$ under MCAR, for $T = 4, 6$ and 1,000 clusters. Distribution (A) satisfies linearity condition (2.11), while distribution (B) has conditionally non-linear moments, violating condition (2.11).

| | | $T = 4$, (A) | | | $T = 4$, (B) | | | $T = 6$, (A) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dropout | | $\hat{\beta}_{GT}$ | | $\hat{\beta}_{WT}$ | $\hat{\beta}_{GT}$ | | $\hat{\beta}_{WT}$ | $\hat{\beta}_{GT}$ | | $\hat{\beta}_{WT}$ |
| Rate | $\rho$ | Indep | Exch | | Indep | Exch | | Indep | Exch | |
| 0.1 | 0 | 100 | 100 | 100 | 96 | 95 | 95 | 100 | 100 | 100 |
| | 0.2 | 98 | 99 | 98 | 95 | 96 | 95 | 101 | 100 | 101 |
| | 0.6 | 85 | 100 | 96 | 86 | 98 | 95 | 78 | 101 | 93 |
| | 0.7 | 78 | 100 | 97 | 76 | 96 | 92 | 65 | 101 | 91 |
| 0.2 | 0 | 101 | 100 | 101 | 90 | 89 | 89 | 100 | 100 | 100 |
| | 0.2 | 96 | 100 | 98 | 89 | 92 | 91 | 99 | 100 | 98 |
| | 0.6 | 82 | 100 | 95 | 71 | 92 | 87 | 68 | 101 | 86 |
| | 0.7 | 65 | 99 | 92 | 58 | 92 | 88 | 55 | 103 | 84 |
| 0.4 | 0 | 101 | 101 | 101 | 73 | 72 | 72 | 103 | 102 | 103 |
| | 0.2 | 97 | 103 | 99 | 77 | 83 | 80 | 105 | 112 | 107 |
| | 0.6 | 71 | 100 | 83 | 51 | 83 | 70 | 60 | 124 | 76 |
| | 0.7 | 56 | 100 | 84 | 56 | 90 | 79 | 40 | 125 | 60 |

Table 2.3: Distribution among 1000 simulation runs of Max($\hat{w}_{it}$), $T = 4$, $K = 1000$ for selected scenarios.

| Dropout | | | | $\hat{\boldsymbol{\beta}}_W$ | | | | | $\hat{\boldsymbol{\beta}}_A$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Type | Rate | $\rho$ | Max($w_{it}$) | Mean | $50^{th}$ | $75^{th}$ | $95^{th}$ | $99^{th}$ | Mean | $50^{th}$ | $75^{th}$ | $95^{th}$ | $99^{th}$ |
| MCAR | 0.2 | 0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.1 | 2.2 | 2.2 | 2.2 | 2.3 | 2.5 | 2.8 |
| | | 0.7 | 2.0 | 2.0 | 2.0 | 2.0 | 2.1 | 2.2 | 2.4 | 2.3 | 2.5 | 2.8 | 3.4 |
| | 0.4 | 0 | 4.7 | 4.9 | 4.9 | 5.1 | 5.5 | 5.9 | 6.2 | 5.9 | 6.6 | 8.5 | 10.7 |
| | | 0.7 | 4.7 | 4.9 | 4.9 | 5.1 | 5.5 | 6.0 | 7.8 | 6.9 | 8.5 | 13.7 | 21.5 |
| MAR | 0.2 | 0 | 2.2 | 2.2 | 2.2 | 2.3 | 2.4 | 2.6 | 2.4 | 2.3 | 2.5 | 2.9 | 3.2 |
| Weak | | 0.7 | 2.2 | 2.2 | 2.2 | 2.3 | 2.4 | 2.5 | 2.4 | 2.4 | 2.5 | 3.0 | 3.7 |
| | 0.4 | 0 | 6.0 | 6.1 | 6.5 | 6.5 | 7.2 | 7.7 | 7.4 | 6.7 | 7.9 | 11.6 | 17.3 |
| | | 0.7 | 6.0 | 6.1 | 6.0 | 6.5 | 7.2 | 8.0 | 8.3 | 7.4 | 8.8 | 14.5 | 22.6 |
| MAR | 0.2 | 0 | 2.8 | 2.8 | 2.8 | 2.9 | 3.1 | 3.2 | 2.9 | 2.8 | 3.2 | 3.8 | 4.4 |
| Strong | | 0.7 | 2.8 | 2.8 | 2.8 | 2.9 | 3.1 | 3.3 | 2.9 | 2.9 | 3.1 | 3.5 | 4.2 |
| | 0.4 | 0 | 9.3 | 9.3 | 9.2 | 10.1 | 11.3 | 12.4 | 11.4 | 9.7 | 12.5 | 21.0 | 36.0 |
| | | 0.7 | 9.3 | 9.4 | 9.3 | 10.2 | 11.6 | 12.4 | 11.5 | 10.4 | 12.5 | 18.9 | 29.4 |

* In all analyses, the minimum $\hat{w}_{it}$ is 1.0.

Table 2.4: Efficiency of $\hat{\beta}_{WT}$ with respect to $\hat{\beta}_{AT}$ under MAR, for $T = 4, 6$ and $K = 1000$, all without auxiliary covariate $\mathbf{V}_i$. Distribution (A) satisfies condition (2.11), while distribution (B) is in violation.

| | | $T = 4$, (A) | | $T = 4$, (B) | | $T = 6$, (A) | |
|---|---|---|---|---|---|---|---|
| Dropout Rate | $\rho$ | MAR Weak | MAR Strong | MAR Weak | MAR Strong | MAR Weak | MAR Strong |
| 0.1 | 0 | 100 | 100 | 96 | 93 | 100 | 100 |
| | 0.2 | 100 | 100 | 93 | 99 | 100 | 96 |
| | 0.6 | 96 | 92 | 92 | 96 | 93 | 92 |
| | 0.7 | 95 | 94 | 92 | 95 | 94 | 92 |
| 0.2 | 0 | 100 | 100 | 91 | 86 | 100 | 101 |
| | 0.2 | 99 | 100 | 91 | 89 | 98 | 96 |
| | 0.6 | 93 | 86 | 88 | 92 | 85 | 81 |
| | 0.7 | 93 | 88 | 89 | 89 | 81 | 73 |
| 0.4 | 0 | 102 | 106 | 78 | 62 | 104 | 106 |
| | 0.2 | 97 | 97 | 81 | 88 | 104 | 108 |
| | 0.6 | 78 | 71 | 73 | 77 | 93 | 70 |
| | 0.7 | 75 | 65 | 73 | 71 | 76 | 77 |

Table 2.5: Estimates of $\beta_6$ ($\times 100$) and estimated standard errors.

| | without $\mathbf{V}_i$ | | | with $\mathbf{V}_i$ | |
|---|---|---|---|---|---|
| Group | $\hat{\beta}_{G6}$ | $\hat{\beta}_{W6}$ | $\hat{\beta}_{A6}$ | $\hat{\beta}_{W6}$ | $\hat{\beta}_{A6}$ |
| Black Males | **-11.0** | **-16.5** | **-14.8** | -12.5 | -10.4 |
| | (7.02) | (7.67) | (6.21) | (7.84) | (6.30) |
| Black Females | **-20.5** | **-22.3** | **-22.1** | **-18.4** | **-19.3** |
| | (5.53) | (6.21) | (5.16) | (6.29) | (5.22) |
| White Males | **-36.9** | **-43.3** | **-39.0** | **-40.3** | **-37.9** |
| | (6.67) | (7.67) | (6.64) | (7.70) | (6.69) |
| White Females | **-53.5** | **-58.2** | **-55.9** | **-54.1** | **-49.9** |
| | (7.04) | (7.97) | (6.87) | (7.99) | (6.81) |

Estimates in bold are significantly different from zero.
Exchangeable correlation was used with $\hat{\beta}_{G6}$; Estimates of $\rho$ range from 0.67 to 0.76.

Table 2.6: Estimates of $\beta_6$ ($\times 100$) and estimated standard errors, by cohort and attained education using the semi-parametric efficient estimator $\hat{\beta}_{A6}$.

| | Birth Cohort 1963 - 1967 | | Birth Cohort 1955 - 1962 | | U.S. Estimate |
|---|---|---|---|---|---|
| | No Degree | College Degree | No Degree | College Degree | |
| Black Males | -0.1 (12.25) | 11.5 (52.42) | -16.0 (7.92) | -24.4 (23.70) | -9.8 (6.73) |
| Black Females | 2.5 (10.56) | -6.7 (27.76) | -30.5 (6.57) | -24.3 (20.32) | -17.6 (5.41) |
| White Males | -22.1 (17.41) | -47.8 (33.70) | -32.6 (8.21) | -66.9 (18.40) | -36.2 (7.41) |
| White Females | -27.9 (15.86) | -39.3 (26.53) | -59.1 (9.51) | -61.6 (16.69) | -49.2 (7.22) |

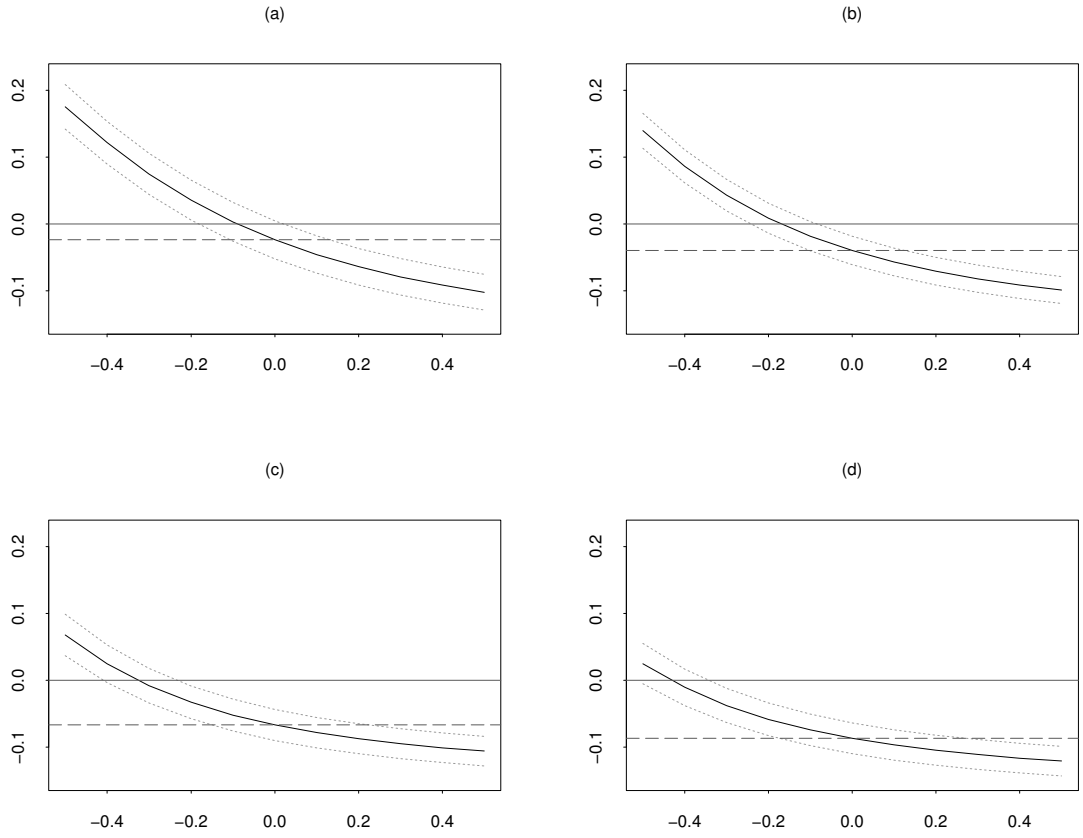


Figure 2.2: The estimated difference between smoking rates at time 6 and smoking rates at time 1, $g_6(\mathbf{X}_i, \hat{\boldsymbol{\beta}}_A) - g_1(\mathbf{X}_i, \hat{\boldsymbol{\beta}}_A)$, with 95% confidence intervals, on the y-axis, as a function of $\alpha_1$ along the x-axis. There is a solid reference line at $\hat{g}_6 - \hat{g}_1 = 0$. The dashed reference line is for $\hat{g}_6 - \hat{g}_1$ estimated when $\alpha_1 = 0$, equivalent to the CARDIA analysis for six time points. When $\alpha_1 \neq 0$, MAR is violated. Graph (a) is for black men, (b) is for black women, (c) is for white men, and (d) is for white women.

# Alternating Logistic Regressions With Improved Finite Sample Properties

## 3.1 Introduction

Associations in correlated binary data are often considered nuisances, however, it is not uncommon that they are scientifically relevant. It may be of interest in this case to model associations more carefully, for example, with covariates defined by traits of clusters or outcome pairs. Accommodating correlation models, estimating equations have been defined for associations characterized by correlations (Prentice, 1988), in addition to estimating equations characterized by odds ratios. Alternating logistic regressions (ALR) was defined to this purpose by Carey, Zeger, and Diggle (1993) to model marginal means of correlated binary outcomes while simultaneously allowing for a flexible association model based on pairwise odds ratios.

Although ALR is useful for making simultaneous inference on marginal mean and association parameters, its use, like the use of estimating equation approaches in general, is subject to concerns regarding their performance in small samples (Emrich and Piedmonte, 1992). The use of estimating equations for association estimation in small samples may result in confidence interval coverage below the nominal level (Evans et al.,

2001; Sharples and Breslow, 1992). Because estimation of correlation parameters in estimating equation procedures are dependent on asymptotic behavior, there may also be bias in their use for small samples (Preisser et al., 2008).

There is limited published information on the finite sample performance of ALR association parameter estimates, however, recent literature suggests that the estimating equations for ALR could be improved with appropriate finite sample adjustments (Preisser et al., 2008). The observed poor performance of estimating equations can be partly attributed to the behavior of sandwich variance estimators, which is consistent while also known to underestimate the actual variance of parameter estimates in small samples (Sharples and Breslow, 1992; Mancl and DeRouen, 2001).

Adjustments for empirical covariance estimators have recently been introduced and have been shown to improve coverage and variance estimation in first order generalized estimating equations (GEE) (Kauermann and Carroll, 2001; Mancl and DeRouen, 2001; Lu et al., 2007). In addition to these methods to improve sandwich variance estimators, small sample bias corrections to the estimating equations of Prentice (1988) were proposed by Sharples and Breslow (1992) and Lu et al. (2007), who both focused on inference for marginal mean parameters. Preisser et al. (2008) were instead concerned with inference for intracluster correlation parameter estimates, demonstrating marked improvement in a bias-adjusted procedure for samples with as few as twenty clusters.

The demonstrated improvement in the estimation and inference for intracluster correlation parameter estimates (Preisser et al., 2008) and the improvement in variance estimation of Mancl and DeRouen (2001) and Kauermann and Carroll (2001) together suggest that there may be utility in an extension of these approaches to ALR. First, by improving the estimation of the variance of association parameter estimates, and second, by reducing the bias of those parameter estimates.

Correcting for the bias in sandwich estimators may not be sufficient to ensure ad-

equate coverage in small samples, as Kauermann and Carroll (2001) showed for independent data. In part because of this, it is not expected that the bias of parameter estimates can be totally eliminated, or that the coverage of confidence intervals reach their nominal level. Despite this, the expectation is that applicability of ALR can be significantly increased by finite sample adjustments to estimating equations and their empirical variance estimates. The aim of this paper is to examine two kinds of finite-sample bias adjustments for ALR, bias corrections in covariance estimators and bias corrections in estimating equations. The impact of these adjustments will be examined on the inference for marginal association model parameters in correlated binary data for which ALR is defined, in simulated data and also in an application to a cluster trial to reduce underage drinking.

## 3.2 Finite sample corrections for ALR

Consider a vector of correlated binary outcomes for subject or cluster $i = 1, \ldots, K$, so that $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ for binary $Y_{ij}$, $1 \leq j \leq n_i$. Let $\mu_{ij} = E(Y_{ij}|\mathbf{X}_i)$ represent the marginal mean of $Y_{ij}$ conditional on covariate $X_i$. The marginal mean $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})'$ is assumed known up to a $p \times 1$ parameter $\boldsymbol{\beta}$. As defined by Liang and Zeger (1986), let $\hat{\boldsymbol{\beta}}$ represent the solution for $\boldsymbol{\beta}$ in the estimating equation

$$\mathbf{U}_{\boldsymbol{\beta}} = \sum_{i=1}^{K} D_i' V_i^{-1} \{ \mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \} = 0 \,, \tag{3.1}$$

where $D_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}'$ and $V_i$ is the working covariance matrix of $\mathbf{Y}_i$. Under mild regulatory conditions, $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ such that $\sqrt{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically multivariate normal with respect to $K$.

In this setting where $\mathbf{Y}_i$ has binary elements, the variance $V_i$ can be characterized

by the pairwise odds ratios

$$\psi_{ijk} \;=\; \frac{P(Y_{ij}=1,Y_{ik}=1)P(Y_{ij}=0,Y_{ik}=0)}{P(Y_{ij}=1,Y_{ik}=0)P(Y_{ij}=0,Y_{ik}=1)}\,,\;\; 1\le j<k\le n_i\,.$$

Assuming that, given a covariate vector $\mathbf{Z}_{ijk}$ for the pair of outcomes $Y_{ij}$ and $Y_{ik}$,

$$\log(\psi_{ijk}) \;=\; \mathbf{Z}'_{ijk}\boldsymbol{\alpha}\,, \tag{3.2}$$

the odds ratio $\psi_{ijk}$ can be modeled through the parameter $\boldsymbol{\alpha}$. Carey et al. (1993) defined alternating logistic regressions (ALR), where $\boldsymbol{\alpha}$ is estimated in a separate estimating equation based on expectations of $Y_{ij}$ conditional on $Y_{ik}$, for $1 \le j < k \le n_i$. The resulting $\hat{\boldsymbol{\alpha}}$ is consistent such that $\sqrt{K}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})$ is asymptotically normal with respect to $K$.

Another procedure for estimating $\boldsymbol{\alpha}$ in (3.2) was defined by Zink and Qaqish (2009), whose $\hat{\boldsymbol{\alpha}}$ is equal to the estimate for $\boldsymbol{\alpha}$ in ALR in a special circumstance. Because the variance estimate for $\hat{\boldsymbol{\alpha}}$ in ALR depends on the ordering of elements in $\mathbf{Y}_i$, a problem resolved by the representation of Zink and Qaqish (2009), the Zink and Qaqish method, known as orthogonalized residuals, will be used here.

Zink and Qaqish (2009) defined a second set of estimating equations for $\boldsymbol{\alpha}$ based on the expectations of cross-products $Y_{ij}Y_{ik}$ conditional on $Y_{ij}$ and $Y_{ik}$. For $\mu_{ijk} = E[Y_{ij}Y_{ik}]$, $\sigma_{ijj} = \mu_{ij}(1-\mu_{ij})$, and $\sigma_{ijk} = \mathrm{cov}(Y_{ij},Y_{ik}) = \mu_{ijk} - \mu_{ij}\mu_{ik}$, orthogonalized residuals estimates $\boldsymbol{\alpha}$ through the residual vector $\mathbf{T}_i$, where $\mathbf{T}_i$ has elements $T_{ijk}$ such that

$$T_{ijk} \;=\; Y_{ij}Y_{ik} - \{\,\mu_{ijk} + b_{ijk:j}(Y_{ij}-\mu_{ij}) + b_{ijk:k}(Y_{ik}-\mu_{ik})\,\}\,,$$

where

$$\begin{aligned} d_{ijk} &= \sigma_{ijj}\sigma_{ikk} - \sigma_{ijk}^2 \\ b_{ijk:j} &= \mu_{ijk}(1-\mu_{ik})(\mu_{ik}-\mu_{ijk})/d_{ijk}, \text{ and} \\ b_{ijk:k} &= \mu_{ijk}(1-\mu_{ij})(\mu_{ij}-\mu_{ijk})/d_{ijk}\,. \end{aligned}$$

In the framework of orthogonalized residuals (Zink and Qaqish, 2009), estimates for $\boldsymbol{\alpha}$ in (3.2) are obtained from the solution to

$$\mathbf{U}_{\boldsymbol{\alpha}} = \sum_{i=1}^{K} S_i' P_i^{-1} \mathbf{T}_i = \mathbf{0}\,. \tag{3.3}$$

The matrix $S_i$ is defined so that $S_i = E\left[-\partial \mathbf{T}_i / \partial \boldsymbol{\alpha}'\right]$ and $P_i$ is an approximate variance of $\mathbf{T}_i$ parameterized with an exchangeable correlation. When this exchangeable correlation is assumed to be zero, the resulting $\hat{\boldsymbol{\alpha}}$ is equivalent to that estimated with ALR (Zink and Qaqish, 2009).

### 3.2.1 Bias-corrected estimating equations

A finite sample correction in the Prentice (1988) framework for estimating the association for correlated binary data was suggested by Preisser et al. (2008), for which there is a related correction to orthogonalized residuals. The uncorrected estimate $\hat{T}_{ijk}$ substituting $\hat{\mu}_{ijk}$, $\hat{\sigma}_{ijj}$, $\hat{\sigma}_{ikk}$, and $\hat{\sigma}_{ijk}$ is consistent but not necessarily unbiased for $T_{ijk}$, thus the solution $\hat{\boldsymbol{\alpha}}$ to the estimating equation (3.3) may also be biased, with bias largest for small samples.

In the Prentice approach, association among elements of $\mathbf{Y}_i$ is characterized by correlations $\rho_{ijk} = \text{corr}(Y_{ij}, Y_{ik})$, as opposed to pairwise odds ratios. Prentice defined estimating equations for correlation parameter $\boldsymbol{\alpha}$ based on the cross-product vector

$\mathbf{R}_i = (R_{i12}, R_{i13}, \ldots, R_{i(n_{i-1})n_i})'$, where $R_{ijk} = r_{ij}r_{ik}$, and $r_{ij} = (Y_{ij} - \mu_{ij})/\sqrt{\sigma_{ijj}}$ (Prentice, 1988). Preisser et al. suggested that $\tilde{\mathbf{R}}_i$ be substituted for $\mathbf{R}_i$ in the estimating equations for $\boldsymbol{\alpha}$, where $\tilde{\mathbf{R}}_i$ has elements $\tilde{R}_{ijk} = \mathbf{G}_{ij.}\hat{\mathbf{R}}_{i.k}$. The vector $\mathbf{G}_{ij.}$ corresponds to the $j^{th}$ row of $G_i = (\, I_{n_i} - H_{1i}\,)^{-1}$ for cluster leverage matrix $H_{1i} = D_i \left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i\right)^{-1} D_i' V_i^{-1}$ (Preisser and Qaqish, 1996) and $\hat{\mathbf{R}}_{i.k} = (\hat{r}_{i1}\hat{r}_{ik}, \ldots, \hat{r}_{in_i}\hat{r}_{ik})'$, where $\hat{r}_{ij} = r_{ij}(\hat{\mu}_{ij})$.

To apply this finite sample bias correction to the framework of orthogonalized residuals, note that $T_{ijk}$, expressed above in terms of cross-products $Y_{ij}Y_{ik}$, can also be expressed in terms of correlations $R_{ijk}$, so that $T_{ijk}$, the elements of vector $\mathbf{T}_i$, are equivalently written

$$T_{ijk} = \sigma_{ijj}^{1/2}\sigma_{ikk}^{1/2}(R_{ijk} - \rho_{ijk}) \; - \; (b_{ijk:j} - \mu_{ik})(Y_{ij} - \mu_{ij}) - (b_{ijk:k} - \mu_{ij})(Y_{ik} - \mu_{ik})\,. \quad (3.4)$$

Let $\tilde{\mathbf{T}}_i$ be an estimate of $\mathbf{T}_i$ in which $\tilde{R}_{ijk}$ is substituted for $R_{ijk}$ in (3.4). Only $\hat{R}_{ijk}$ is corrected; correcting $(Y_{ij} - \hat{\mu}_{ij})$ and $(Y_{ik} - \hat{\mu}_{ik})$ in $\hat{\mathbf{T}}_i$ has negligible effect. The vector $\tilde{\mathbf{T}}_i$ can be substituted for estimate $\hat{\mathbf{T}}_i$, with elements $\hat{T}_{ijk}$, in (3.3) for less biased estimation of $\boldsymbol{\alpha}$ when the number of clusters is small.

The finite sample corrected estimating equations of Preisser et al. (2008) are based on a Taylor series expansion of residual $\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i$ around $\boldsymbol{\beta}$. This finite sample correction substituting $\tilde{R}_{ijk} = \mathbf{G}_{ij.}\hat{\mathbf{R}}_{i.k}$ will be referenced as matrix multiplicative adjusted estimating equations (MMEE).

### 3.2.2 Bias-corrected covariance estimation

Letting $\boldsymbol{\Omega} = \left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i\right)$, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is consistently estimated by

$$\boldsymbol{\Omega}^{-1} \left( \sum_{i=1}^{K} D_i' V_i^{-1} B_{1i} \left\{ \mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \right\} \left\{ \mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \right\}' B_{1i}' V_i^{-1} D_i \right) \boldsymbol{\Omega}^{-1}\,. \quad (3.5)$$

The covariance matrix of $\hat{\boldsymbol{\alpha}}$ is likewise consistently estimated by

$$\left(\sum_{i=1}^{K} S_i' P_i^{-1} S_i\right)^{-1} \left(\sum_{i=1}^{K} S_i' P_i^{-1} B_{2i}\, \mathbf{T}_i\, \mathbf{T}_i'\, B_{2i}' P_i^{-1} S_i\right) \left(\sum_{i=1}^{K} S_i' P_i^{-1} S_i\right)^{-1} . \qquad (3.6)$$

Substituting $B_{1i} = I_{n_i}$ in (3.5) and $B_{2i} = I_{m_i}$ in (3.6) gives the sandwich estimators for the covariance of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ in their standard form, where $m_i$ is the number of observation pairs in cluster $i$. These variance estimators will henceforth be referred to as BC0. Substituting $B_{1i} = (I_{n_i} - H_{1i})^{-1}$ in (3.5) for $H_{1i} = D_i \left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i\right)^{-1} D_i' V_i^{-1}$ yields the sandwich estimator for $\text{cov}(\hat{\boldsymbol{\beta}})$ described by Mancl and DeRouen (2001). Substituting $B_{2i} = (I_{m_i} - H_{2i})^{-1}$ in (3.6), where $H_{2i} = S_i \left(\sum_{i=1}^{K} S_i' P_i^{-1} S_i\right)^{-1} S_i' P_i^{-1}$, yields the sandwich estimator for $\text{cov}(\hat{\boldsymbol{\alpha}})$ analogous to Mancl and DeRouen's estimator for $\text{cov}(\hat{\boldsymbol{\beta}})$. The variance estimators with Mancl and DeRouen's adjustment will be referred to as BC2, following the notation used by Lu et al. (2007). Both BC0 and bias-corrected BC2 will be used to estimate the variance of $\hat{\boldsymbol{\alpha}}$ in the analysis of simulated and actual data.

## 3.3 Simulation

Clustered binary random variates will be generated and analyzed to determine the advantage of estimating association parameter $\boldsymbol{\alpha}$ with finite sample adjusted ALR compared to standard ALR. Taking the analysis of the underage drinking analysis as a starting point, each realization will have $K = 20$, 40, 80, or 120 clusters with $n = 30$ observations each. Unlike the data used later in an application, the number of observations across clusters will be held constant. Variates $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in})'$ will be generated with mean $\mu_{ij} = E(Y_{ij}|X_{1ij}, X_{2ij})$, $1 \leq j \leq n$, such that

$$\text{logit}(\mu_{ij}) \;=\; \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{1ij} X_{2ij} \, . \qquad (3.7)$$

This marginal mean represents the prevalence of an outcome for two groups at two times, where $X_{1ij} = 1$ for an intervention community and 0 otherwise, and $X_{2ij}$ is an indicator for posttest. The parameter $\beta_0$ is the log prevalence of the last 30-day alcohol use in the control group at baseline. The parameter $\beta_1$ represents an initial difference between intervention and control communities, while the parameter $\beta_3$ represents the difference in effect between intervention and control communities over time.

Parameters $(\beta_1, \beta_2, \beta_3)'$ will be fixed at $(0, -0.10, -0.25)'$ and $\beta_0$ will be varied within $(-0.5, 0.25)'$. These $\mathbf{Y}_i$ will be generated with marginal means (3.7) such that their association is restricted by

$$\log(\psi_{ijk}) \quad = \quad \alpha_1 Z_{1ijk} + \alpha_2 Z_{2ijk} \, . \tag{3.8}$$

The covariate vector $\mathbf{Z}_{ijk} = (Z_{1ijk}, Z_{2ijk})'$ is either $(1, 0)'$ when $X_{2ij} = X_{2ik}$, so that $\alpha_1$ represents the log odds ratio within time, or $\mathbf{Z}_{ijk} = (0, 1)'$ when $X_{2ij} \neq X_{2ik}$, so that $\alpha_2$ represents the log odds ratio between times. The parameter $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$ is varied within $\{(0.1, 0.05)', (0.05, 0.025)'\}$, as associations in cluster trials tend to be small and to degrade over time.

All data is generated with the algorithm of Qaqish (2003) after converting odds ratios to correlations (Mardia, 1967; Preisser et al., 2002). In the analysis of data simulated with (3.7) and (3.8), each realization of simulated data will be analyzed using standard ALR and re-analyzed using the proposed bias-corrected ALR. Standard errors in both cases will be estimated using standard sandwich estimators and the bias-corrected sandwich estimator, so that the bias of $\hat{\boldsymbol{\alpha}}$ and the coverage of nominally 95% confidence intervals will be examined.

### 3.3.1 Simulation results

Bias of $\hat{\boldsymbol{\alpha}}$ in both standard ALR and MMEE estimates is in Table 3.1. The $\hat{\alpha}_1$ in standard ALR always underestimates $\alpha_1$, by as much as two thirds when $K = 20$. Even for $K = 80$, $\alpha_1$ is underestimated by standard ALR by as much as 15%. The bias of $\hat{\alpha}_1$ in MMEE adjusted ALR is generally negative, and between 1% and 8% at $K = 20$, reducing somewhat for larger $K$. The absolute bias for $\hat{\alpha}_1$ in MMEE adjusted ALR is always less than that of standard ALR.

The between time association estimate $\hat{\alpha}_2$ was also generally underestimated by standard ALR, although relatively by less than for $\hat{\alpha}_1$. At $K = 20$ clusters and $\beta_0 = -0.5$, the bias of $\hat{\alpha}_2$ in standard ALR outperforms that of MMEE adjusted ALR, which overestimates $\alpha_2$. When $K = 80$ clusters, $\alpha_2$ is mostly underestimated by standard ALR, by as much as 7%, while the bias of $\hat{\alpha}_2$ in MMEE adjusted ALR is close to 5%. Bias in MMEE adjusted ALR for $\hat{\alpha}_2$ compared to $\hat{\alpha}_1$ is relatively constant across $K$, with greater relative magnitude.

The relative bias of variance estimators BC0 and BC2 is shown in Table 3.2 for both $\alpha_1$ and $\alpha_2$ association parameter variance estimates. The average of BC0 and BC2 across simulations is compared to the Monte Carlo variance of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ in order to estimate bias. Note that in general relative bias declines with larger $K$, and that variance estimator BC2 is less biased than BC0 for both standard ALR and MMEE adjusted ALR. Note also that the relative performance of BC0 and BC2 is consistent across estimating equation standard or adjusted ALR.

Coverage of uncorrected and bias-corrected nominally 95% confidence intervals for $\hat{\boldsymbol{\alpha}}$ in both standard ALR and MMEE is shown in Table 3.3. Confidence interval coverage for the case that $\beta_0 = -0.5$ and $\boldsymbol{\alpha} = (0.05, 0.025)'$ is also shown in Figure 3.1. It is notable first of all that for $\hat{\alpha}_1$, even at $K = 80$ clusters the standard ALR procedure without variance adjustment as it is commonly implemented is below nominal coverage

rates, staying close between 91% and 92%. The coverage of MMEE adjusted ALR is larger in every case, from one to two percentage points, and never exceeding the nominal level.

In fact for the within time association $\alpha_1$, the coverage for MMEE always exceeds that of standard ALR, by a margin upwards of 5% in some cases. When $K = 20$, the coverage of standard ALR confidence intervals is very low both with and without variance estimator adjustment BC2. Note that variance adjustment BC2 always increases coverage (Preisser et al., 2008).

For the between time association $\alpha_2$, the coverage of standard ALR confidence intervals is markedly improved. There is less contrast between the coverage rates standard ALR and MMEE in this case than for $\hat{\alpha}_1$, however, standard ALR is still generally outperformed by the confidence intervals of MMEE adjusted ALR. The coverage for $\hat{\alpha}_2$ is very close to the nominal level of 95% at $K = 80$ for all examined methods, and generally only a few points below nominal at $K = 40$. In some cases coverage negligibly exceeds the nominal level, especially for $K = 120$.

## 3.4 Example

The Enforcing Underage Drinking Laws (EUDL) Program is a nonrandomized community intervention trial begun in 1998 as part of a federal initiative to reduce underage drinking (Wolfson et al., 2004). The intervention was evaluated with a nested cross sectional design based on repeated random telephone surveys of individuals aged 16 to 20 years from 202 communities, half of which were comparison communities matched by a propensity score based on U.S. Census data. The primary outcomes were binary, including self-reported last 30-day alcohol use.

A subset of the data is analyzed consisting of baseline and year one follow-up data from 38 communities. Cluster sizes range from 27 to 41 with mean size 35.4. The goal

of this analysis is to estimate the pairwise odds ratio of the binary outcome self-reported last 30-day alcohol use, while also assessing the effect of the intervention.

Let $Y_{ij}$ be a binary indicator of last-30 day alcohol use for subject $j = 1, \ldots, n_i$ in community $i$. The model for marginal mean $\mu_{ij} = E[Y_{ij} \,|\mathbf{X}_{ij}]$ for $\mathbf{X}_{ij} = (X_{1ij}, X_{2ij}, X_{3ij}, \mathbf{X}_{ij}^{*\prime})'$ is given by

$$\text{logit}(\mu_{ij}) \quad = \quad \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \mathbf{X}_{ij}^{*'}\boldsymbol{\beta}^* \,, \tag{3.9}$$

where $X_{1ij} = 1$ for an intervention community and 0 for a control community, $X_{2ij} = 1$ for posttest and 0 for baseline, $X_{3ij} = X_{1ij}X_{2ij}$, and $\mathbf{X}_{ij}^*$ is a vector of covariates not related to time or intervention. The pairwise odds ratio $\psi_{ijk}$ for outcomes $Y_{ij}$ and $Y_{ik}$ in community $i$ is modeled with (3.8) where $(Z_{1ijk}, Z_{2ijk})' = (0, 1)'$ when $X_{2ij} = X_{2ik}$, and $(Z_{1ijk}, Z_{2ijk})' = (1, 0)'$ when $X_{2ij} \neq X_{2ik}$. The first pairwise odds ratio ($\exp\{\alpha_1\}$) is among individuals in the same community at a specific time point, and the second ($\exp\{\alpha_2\}$) is among individuals in the same community at different time points. Data from the EUDL Trial is modeled here with (3.8) and (3.9) using both the estimating equations of standard ALR and the estimating equations with a finite sample correction defined in Section 3.2. The parameter estimates from the model defined by (3.8) and (3.9) and their standard errors as estimated by both BC0 and BC2 are in Table 3.4.

Note first that there is little discernable difference in marginal mean parameter estimates between standard ALR and MMEE. Because the fitting algorithm of these procedures share an estimating equation for $\boldsymbol{\beta}$, this result is not unexpected. The difference in the approaches of standard ALR and MMEE are their methods of estimating $\boldsymbol{\alpha}$ in the usual context of generalized estimating equations, and so any differences in $\hat{\boldsymbol{\beta}}$ would be attributed to differences in $\hat{\boldsymbol{\alpha}}$. The intervention effect represented by $\beta_3$ in model (3.9) is non-significant across all standard ALR and MMEE estimates, as was expected from the results of previous analyses (Preisser et al., 2008).

The BC2 estimator provides larger standard error estimates than BC0, applying to both the marginal mean model and the association model for pairwise odds ratios. Because residual vectors are typically underestimated and the robust variance estimator of Liang and Zeger (1986) is known to underestimate the true variance of $\hat{\boldsymbol{\beta}}$ (Mancl and DeRouen, 2001), this is in agreement with expectations.

Note also the parameter estimates in Table 3.4 for the pairwise odds ratio model. MMEE estimates of within-time and between-time $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are notably larger than their corresponding standard ALR estimates, by 33 % and 35%, respectively. This reflects the reduction of bias in $\hat{\boldsymbol{\alpha}}$ obtained with MMEE, as summarized by simulated data in Table 3.1, and the corresponding observed tendency of standard ALR to underestimate association parameters.

Although the standard error estimates depend on the estimation method used, BC0 or BC2, these differences in the EUDL analysis are not as marked as those determined by the estimating equation method, standard ALR or MMEE. Point estimates for the odds ratios corresponding to these $\hat{\boldsymbol{\alpha}}$ with their 95% confidence intervals, both for BC0 and BC2, are shown in Figure 3.2.

In addition to these results, Table 3.5 is included with $\hat{\boldsymbol{\alpha}}$ and estimated standard errors in the equivalent association model (3.8) for nine different dichotomous outcomes collected in the EUDL study. The $\hat{\boldsymbol{\alpha}}$ in these models varies widely across standard ALR and ALR with the proposed MMEE adjustment, although in general $\hat{\boldsymbol{\alpha}}$ is not significantly different from zero. Each outcome is described briefly in Table 3.5, and a detailed description is available in Preisser et al. (2007). The mean for each outcome is modeled with (3.9) as in the above analysis, and also using a model with only an intercept.

## 3.5 Discussion

Data was simulated so that the circumstances of analysis would resemble a cluster survey repeated over time, where cluster sizes are large and associations within clusters are small. Although the variations considered in simulated data may fairly represent the EUDL trial, there are many circumstances where alternating logistic regressions are often used that cannot be represented or interpolated from the results here. Large associations in particular were not considered, and neither were data with small clusters or with varying cluster sizes. Large clusters present a particular challenge from the standpoint of analysis, however, and the adjustments proposed here were shown in the simulation study to be useful in that they estimate association parameters with less bias than standard ALR, consistently across $K$ and for varying marginal means.

Although the proposed adjustment to the estimating equations of ALR did not always provide association parameter estimates with less bias than standard ALR, the reduction in bias was often marked. In addition, the variance estimators in the adjusted estimating equations were less biased than those of standard ALR. Variance estimates were additionally improved by using an extension of the bias-corrected variance estimator introduced by Mancl and DeRouen (2001).

These two effects (reduced bias of association parameter estimates, improved variance estimation) both contributed to the coverage rates of nominally 95% confidence intervals. The coverage of 95% confidence intervals in standard ALR for simulated data was observed approaching 80%; in this scenario the adjusted estimating equations coverage, although still below nominal levels, was improved to approximately 90%.

Results from these analyses with simulated data indicate that the proposed small sample adjusted ALR is an appropriate analysis for the chosen subset of underage drinking data, and possibly for other repeated cluster survey data. Overall this paper has shown that ALR can be applied with increased accuracy to small samples, and that

the proposed method improves inference for association parameters in repeated binary data when applying alternating logistic regressions.

Table 3.1: Estimated bias, ( Average $\{\hat{\alpha}-\alpha\}$ ), in alternating logistic regressions (ALR) for standard ALR and ALR with a multiplicative matrix adjustment using $\tilde{R}_{ijk} = \hat{G}_i[j,]\hat{C}_i[,k]$ (MMEE). These results are for 1000 simulations of 40, 80, or 120 clusters each, with cluster size $n = 30$.

| | | Bias $\hat{\alpha}_1$ | | Bias $\hat{\alpha}_2$ | |
|---|---|---|---|---|---|
| $\beta_0$ | $\boldsymbol{\alpha}$ | ALR | MMEE | ALR | MMEE |
| | | $K = 20$ | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | -.0424 | -.0040 | -.0007 | .0047 |
| | $(0.05, 0.025)'$ | -.0323 | .0022 | -.0002 | .0025 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | -.0354 | .0015 | -.0059 | -.0010 |
| | $(0.05, 0.025)'$ | -.0357 | -.0040 | -.0049 | -.0026 |
| | | $K = 40$ | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | -.0198 | -.0006 | .0012 | .0039 |
| | $(0.05, 0.025)'$ | -.0146 | .0026 | -.0008 | .0005 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | -.0185 | -.0002 | -.0037 | -.0012 |
| | $(0.05, 0.025)'$ | -.0179 | -.0021 | -.0020 | -.0008 |
| | | $K = 80$ | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | -.0102 | -.0006 | .0014 | .0027 |
| | $(0.05, 0.025)'$ | -.0067 | .0018 | -.0018 | -.0012 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | -.0085 | .0006 | -.0003 | .0009 |
| | $(0.05, 0.025)'$ | -.0086 | -.0007 | -.0016 | -.0010 |
| | | $K = 120$ | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | -.0060 | .0004 | .0013 | .0022 |
| | $(0.05, 0.025)'$ | -.0056 | -.0000 | -.0011 | -.0007 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | -.0061 | -.0000 | -.0015 | -.0006 |
| | $(0.05, 0.025)'$ | -.0051 | .0002 | -.0005 | -.0001 |

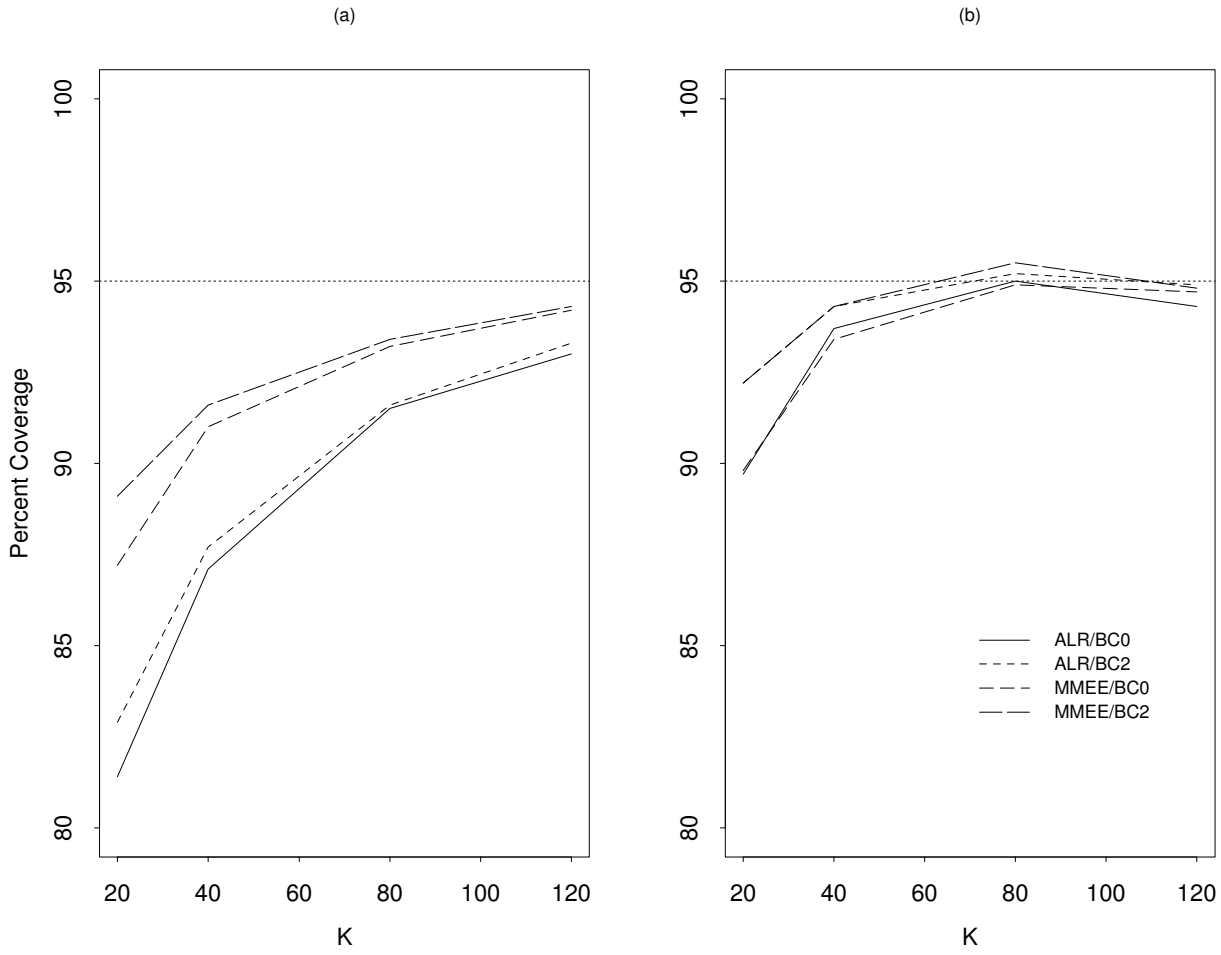


Figure 3.1: Coverage of nominally 95% confidence intervals for standard ALR and ALR with proposed MMEE adjustment, using both BC0 and BC2 variance estimates. Plot (a) is for $\hat{\alpha}_1$ (within time) and plot (b) is for $\hat{\alpha}_2$ (between time).

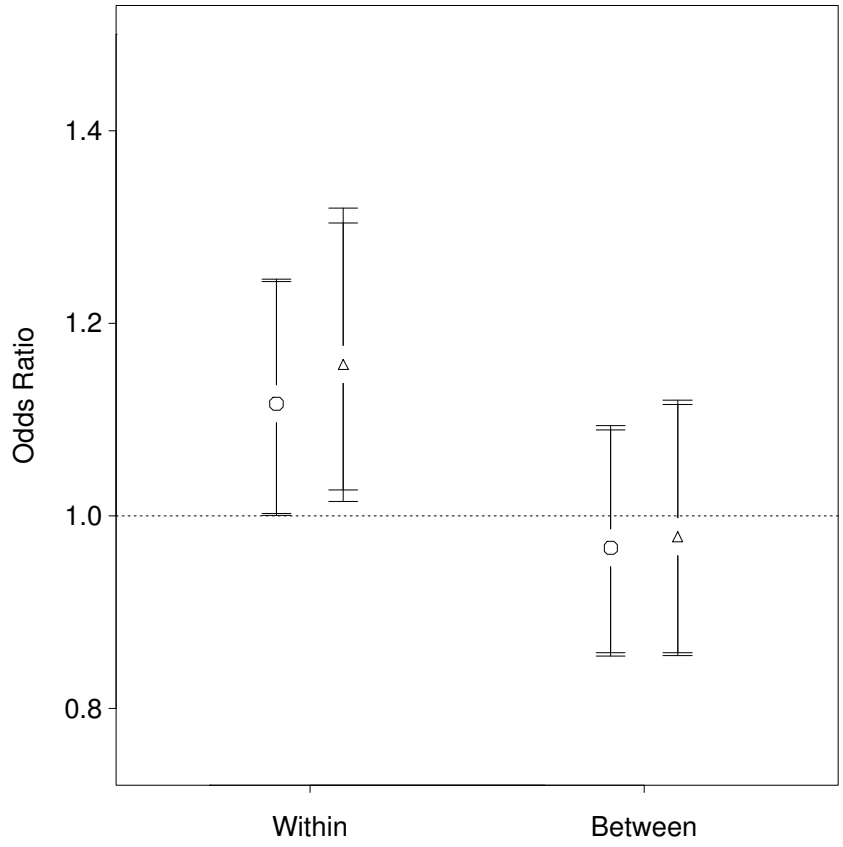


Figure 3.2: Odds Ratio estimates of within and between time associations in the EUDL data. The estimates marked with a circle are those of standard ALR, and those marked with a triangle are estimated with ALR and a finite sample correction. Both BC0 and BC2 are represented in 95% confidence intervals; BC0 yields the smaller interval in all cases.

Table 3.2: Estimated percent relative bias in standard alternating logistic regressions (ALR) and in ALR with proposed MMEE adjustment, of variance estimators BC0 and BC2. These results are for 1000 simulations of 40, 80, or 120 clusters each, with cluster size $n = 30$. Bias is measured relative to the Monte Carlo variance of $\hat{\boldsymbol{\alpha}}$.

| | | Bias $\hat{\text{var}}(\hat{\alpha}_1)$ | | | | Bias $\hat{\text{var}}(\hat{\alpha}_2)$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | ALR | | MMEE | | ALR | | MMEE | |
| $\beta_0$ | $\boldsymbol{\alpha}$ | BC0 | BC2 | BC0 | BC2 | BC0 | BC2 | BC0 | BC2 |
| | | | | $K = 20$ | | | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | -10.6 | -0.8 | -10.3 | -0.5 | -18.6 | -9.7 | -18.6 | -9.7 |
| | $(0.05, 0.025)'$ | -20.5 | -11.9 | -20.5 | -11.8 | -18.4 | -9.5 | -18.4 | -9.5 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | -14.2 | -4.9 | -14.3 | -5.0 | -16.3 | -7.3 | -16.3 | -7.3 |
| | $(0.05, 0.025)'$ | -14.5 | -5.2 | -14.5 | -5.2 | -17.1 | -8.2 | -17.1 | -8.1 |
| | | | | $K = 40$ | | | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | -8.1 | -3.3 | -7.9 | -3.1 | -1.5 | 3.6 | -1.5 | 3.7 |
| | $(0.05, 0.025)'$ | -13.2 | -8.7 | -13.2 | -8.7 | -5.0 | -0.1 | -5.0 | 0.0 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | -11.8 | -7.3 | -12.0 | -7.5 | -12.3 | -7.7 | -12.3 | -7.7 |
| | $(0.05, 0.025)'$ | -10.0 | -5.3 | -10.0 | -5.4 | -4.3 | 0.7 | -4.3 | 0.7 |
| | | | | $K = 80$ | | | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | 1.6 | 4.2 | 1.9 | 4.5 | -0.3 | 2.2 | -0.3 | 2.3 |
| | $(0.05, 0.025)'$ | -8.5 | -8.1 | -8.4 | -6.1 | 3.4 | 6.1 | 3.4 | 6.1 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | -6.7 | -4.4 | -6.7 | -4.3 | -6.7 | -4.3 | -6.7 | -4.3 |
| | $(0.05, 0.025)'$ | -9.3 | -7.0 | -9.4 | -7.1 | -6.7 | -4.3 | -6.7 | -4.3 |
| | | | | $K = 120$ | | | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | 1.6 | 3.3 | 1.7 | 3.5 | 7.5 | 9.3 | 7.6 | 9.4 |
| | $(0.05, 0.025)'$ | -2.6 | -0.9 | -2.5 | -0.8 | -3.8 | -2.1 | -3.7 | -2.0 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | -5.1 | -3.5 | -5.0 | -3.4 | -0.6 | 1.1 | -0.5 | 1.2 |
| | $(0.05, 0.025)'$ | -8.7 | -7.1 | -8.7 | -7.2 | -0.5 | 1.1 | -0.5 | 1.2 |

Table 3.3: Coverage of nominal 95% confidence intervals in alternating logistic regressions (ALR) for standard ALR and ALR with a multiplicative matrix adjustment using $\tilde{R}_{ijk} = \hat{G}_i[j,]\hat{C}_i[,k]$ (MMEE), for both BC0 and BC2 standard error estimators. These results are for 1000 simulations, with cluster size $n = 30$.

| | | Coverage $\hat{\alpha}_1$ | | | | Coverage $\hat{\alpha}_2$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | ALR | | MMEE | | ALR | | MMEE | |
| $\beta_0$ | $\boldsymbol{\alpha}$ | BC0 | BC2 | BC0 | BC2 | BC0 | BC2 | BC0 | BC2 |
| | | | | $K = 20$ | | | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | 80.6 | 82.4 | 89.2 | 90.6 | 90.9 | 92.4 | 91.3 | 93.1 |
| | $(0.05, 0.025)'$ | 81.4 | 82.9 | 87.2 | 89.1 | 89.7 | 92.2 | 89.8 | 92.2 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | 83.0 | 83.8 | 88.6 | 90.1 | 90.8 | 92.5 | 91.2 | 92.8 |
| | $(0.05, 0.025)'$ | 80.4 | 82.2 | 88.7 | 90.4 | 91.2 | 93.1 | 91.7 | 93.5 |
| | | | | $K = 40$ | | | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | 87.6 | 88.3 | 91.7 | 92.7 | 94.7 | 95.7 | 95.1 | 95.5 |
| | $(0.05, 0.025)'$ | 87.1 | 87.7 | 91.0 | 91.6 | 93.7 | 94.3 | 93.4 | 94.3 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | 88.8 | 90.1 | 92.3 | 92.5 | 91.8 | 92.4 | 92.2 | 93.1 |
| | $(0.05, 0.025)'$ | 86.2 | 87.6 | 90.9 | 91.5 | 93.7 | 94.2 | 93.8 | 94.5 |
| | | | | $K = 80$ | | | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | 92.4 | 92.5 | 94.1 | 94.3 | 95.2 | 95.5 | 95.2 | 95.4 |
| | $(0.05, 0.025)'$ | 91.5 | 91.6 | 93.2 | 93.4 | 95.0 | 95.2 | 94.9 | 95.5 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | 91.3 | 92.1 | 92.9 | 92.9 | 93.6 | 93.8 | 93.5 | 94.2 |
| | $(0.05, 0.025)'$ | 90.8 | 90.9 | 92.5 | 92.9 | 93.4 | 94.1 | 93.6 | 93.8 |
| | | | | $K = 120$ | | | | | |
| -0.50 (0.38) | $(0.1, 0.05)'$ | 93.1 | 93.1 | 94.6 | 94.9 | 95.2 | 95.4 | 95.2 | 95.3 |
| | $(0.05, 0.025)'$ | 93.0 | 93.3 | 94.2 | 94.3 | 94.3 | 94.9 | 94.7 | 94.8 |
| 0.25 (0.56) | $(0.1, 0.05)'$ | 92.7 | 92.7 | 94.3 | 94.4 | 94.3 | 94.5 | 94.5 | 94.6 |
| | $(0.05, 0.025)'$ | 92.3 | 92.8 | 93.7 | 94.0 | 94.2 | 94.6 | 94.1 | 94.3 |

Table 3.4: Parameter estimates and their standard errors for self-reported last 30-day alcohol use among youth in the community trial to reduce underage drinking (EUDL) based upon the uncorrected sandwich estimator (BC0) and the bias-corrected (BC2) covariance estimator comparing uncorrected estimating equations (ALR) and the adjusted estimating equations (MMEE)

| | ALR | | | MMEE | | |
|---|---|---|---|---|---|---|
| Parameter | Est. | BC0 | BC2 | Est. | BC0 | BC2 |
| *Marginal mean model* | | | | | | |
| Intercept ($\beta_0$) | -0.576 | 0.152 | 0.164 | -0.576 | 0.152 | 0.164 |
| Intervention ($\beta_1$) | 0.029 | 0.205 | 0.225 | 0.030 | 0.205 | 0.225 |
| Posttest ($\beta_2$) | -0.085 | 0.156 | 0.165 | -0.084 | 0.157 | 0.165 |
| Intvn $\times$ Post ($\beta_3$) | -0.018 | 0.281 | 0.298 | -0.022 | 0.281 | 0.298 |
| Male gender | 0.103 | 0.116 | 0.116 | 0.121 | 0.115 | 0.119 |
| Age = 18 | 0.619 | 0.126 | 0.131 | 0.624 | 0.126 | 0.131 |
| Age = 19/20 | 1.206 | 0.137 | 0.141 | 1.204 | 0.137 | 0.141 |
| Michigan | -0.143 | 0.125 | 0.137 | -0.144 | 0.125 | 0.138 |
| Ohio | -0.662 | 0.262 | 0.326 | -0.664 | 0.261 | 0.323 |
| *Pairwise Odds Ratio model* | | | | | | |
| Within ($\alpha_1$) | 0.110 | 0.055 | 0.056 | 0.146 | 0.061 | 0.063 |
| Between ($\alpha_2$) | -0.034 | 0.061 | 0.063 | -0.022 | 0.067 | 0.069 |

Table 3.5: Association parameter estimates in the model (3.8) for the pairwise odds ratio for different dichotomous outcomes in the EUDL data, when the full model for the marginal mean is used as specified by (3.9), and a model with only an intercept, i.e., $\text{logit}(\mu_{ij}) = \beta_0$.

| | | Full Model | | Intercept Model | |
|---|---|---|---|---|---|
| Measure (Prevalence) | $\alpha$ | ALR | MMEE | ALR | MMEE |
| Binge drinking (0.1980) | | | | | |
| | Within | 0.04 (.059) | 0.09 (.070) | 0.05 (.062) | 0.06 (.070) |
| | Between | 0.03 (.065) | 0.05 (.074) | 0.06 (.060) | 0.07 (.064) |
| DWI drive (0.0578) | | | | | |
| | Within | -0.25 (.113) | -0.13 (.128) | -0.12 (.140) | -0.10 (.155) |
| | Between | -0.04 (.117) | -0.01 (.133) | 0.02 (.123) | 0.04 (.130) |
| Past 30-day alcohol use (0.4212) | | | | | |
| | Within | 0.11 (.050) | 0.15 (.061) | 0.14 (.062) | 0.15 (.065) |
| | Between | -0.03 (.058) | -0.02 (.067) | 0.01 (.066) | 0.02 (.067) |
| Past 7-day alcohol use (0.2459) | | | | | |
| | Within | 0.03 (.060) | 0.07 (.068) | 0.07 (.070) | 0.07 (.074) |
| | Between | 0.02 (.046) | 0.03 (.054) | 0.05 (.056) | 0.05 (.058) |
| Attempt to purchase alcohol (0.0536) | | | | | |
| | Within | 0.07 (.207) | 0.23 (.279) | 0.15 (.186) | 0.17 (.209) |
| | Between | -0.01 (.182) | 0.05 (.220) | 0.10 (.183) | 0.12 (.200) |
| Nonviolent consequences to alcohol use (0.3658) | | | | | |
| | Within | 0.02 (.040) | 0.04 (.045) | 0.05 (.046) | 0.05 (.049) |
| | Between | 0.03 (.037) | 0.04 (.046) | 0.04 (.043) | 0.04 (.048) |
| Perception of alcohol use among peers (0.55242) | | | | | |
| | Within | 0.03 (.045) | 0.06 (.046) | 0.04 (.046) | 0.04 (.045) |
| | Between | -0.01 (.046) | -0.00 (.047) | -0.02 (.049) | -0.01 (.046) |
| Perception of getting caught by police (0.3990) | | | | | |
| | Within | 0.01 (.040) | 0.04 (.046) | 0.04 (.051) | 0.05 (.052) |
| | Between | 0.03 (.037) | 0.04 (.041) | 0.05 (.046) | 0.05 (.047) |
| Commercial source of alcohol (0.0720) | | | | | |
| | Within | -0.15 (.123) | -0.03 (.140) | -0.03 (.129) | -0.01 (.135) |
| | Between | 0.13 (.122) | 0.17 (.144) | 0.26 (.139) | 0.28 (.152) |

# Alternating Logistic Regressions for Ordinal Data

## 4.1 Introduction

Methodology for the analysis of multivariate data is currently a very active area in the statistical literature. Although continuous response models have received a lot of this attention, correlated categorical responses also arise in various biomedical applications. Ordinal response models provide a generalization to multilevel outcomes for methods available specifically for correlated binary data.

There are a number of methods for modeling correlated ordinal or multinomial data that have been proposed in the statistical literature. These methods include those based on marginal methods (Prentice and Zhao, 1991; Liang et al., 1992; Heagerty and Zeger, 1996) and subject specific heirarchical models (Ezzet and Whitehead, 1991; Agresti and Lang, 1993; Crouchley, 1995). Likelihood-based methods for correlated multinomial or ordinal data were proposed by Dale (1986), who introduced a likelihood-based model for bivariate ordinal data using the Plackett distribution, and whose method was extended to multivariate multinomial data by Molenberghs and Lesaffre (1994). A likelihood-based probit model for correlated multinomial data was proposed by Lesaffre and Molenberghs (1991), and Glonek and McCullagh (1995) introduced an alternate class of models for correlated multinomial data using the multivariate logistic trans-

form of McCullagh and Nelder (1989). The method proposed by Glonek and McCullagh (1995) analyzes the dependency of the joint distribution of multinomial outcomes on covariates.

In contrast to maximum likelihood methods, estimating equations for marginal models do not employ the full joint distributions of multinomial outcomes to estimate model parameters. A certain class of marginal model has come into wide use for correlated data with the advent of generalized estimating equations, introduced by Liang and Zeger (1986).

The application of generalized estimating equations to ordinal or categorical data has received considerable attention in the statistical literature to date. Liang, Zeger and Qaqish (1992) defined a marginal model for multinomial data using generalized estimating equations based on response vectors and vectors of response cross-products. Lipsitz et al. (1994) also proposed estimating equations for clustered categorical data based on the estimating equations of Liang and Zeger (1986). Lipsitz et al. (1994) outlined the iterative estimation of the covariance for a select number of structures (exchangeable, 1-dependence, banded, and unstructured) based on a method of moments approach.

Marginal methods based on generalized estimating equations for ordinal data have also been examined by Clayton (1992), particularly in comparison to maximum likelihood. Gange, Linton, Scott, et al. (1995) applied generalized estimating equations to bivariate ordinal data, and Miller, Davis, and Landis (1993) showed that under certain assumptions generalized estimating equations estimators are equal to those of weighted least squares.

Also based on generalized estimating equations (Liang and Zeger, 1986) alternating logistic regressions (ALR) was introduced by Carey, Zeger, and Diggle (1993) in the analysis of multivariate binary data. Heagerty and Zeger (1996) defined estimating

equations for the associations of correlated multinomial data in an adaptation of ALR to ordinal data. Heagerty and Zeger (1996) compared the efficiency for estimating association parameters with ALR and different marginal methods for categorical data, including second order estimating equations (Liang et al., 1992).

Second order estimating equations for ordinal data solve simultaneously for mean and association parameters. This method can be computationally burdensome for large clusters, having a matrix of dimension $Cn + C^2\binom{n}{2}$ to invert, where $C + 1$ is the number of multinomial response levels, and $n$ is the cluster size. First order generalized estimating equations are less burdensome computationally for large clusters, and have high efficiency for association parameters when outcome correlation is not large.

There are certain deficiencies in the ALR Heagerty and Zeger (1996) defined for correlated multinomial outcomes, as there are in the ALR defined for correlated binary outcomes. In particular, although the Heagerty and Zeger (1996) estimate of association parameter $\boldsymbol{\alpha}$ is invariant to the order of observations within cluster, the robust estimate of var($\hat{\boldsymbol{\alpha}}$) is not. In addition to this drawback, because the derivative matrix is stochastic, standard estimating equation theory is not applicable.

Zink and Qaqish (2009) defined estimating equations for association parameters in binary data which yield parameter estimates equal to the ALR $\hat{\boldsymbol{\alpha}}$ in special case, while resolving the disadvantages of estimating $\boldsymbol{\alpha}$ with ALR. The aim of this research is to extend the method of Zink and Qaqish (2009) to multinomial outcomes. This methodology is defined in detail in Sections 2 and 3. An analysis with the orthogonalized residual methodology using data from a study of post-operative altered sensation is described in Section 4. A small simulation study is described in Section 5.

## 4.2 Alternating logistic regressions for ordinal data

Let $O_{ij}$ be ordinal measurement $j$ in cluster $i$, for $i = 1, \ldots, K$, where cluster $i$ has $n_i$ observations. This measurement has $C + 1$ levels or possible realizations, so that $O_{ij} = c$ for some $c \in 1, \ldots, C + 1$. Now let $\mathbf{Y}_{ij}$ be a vector representation of $O_{ij}$ with binary elements

$$Y_{ij}^{(c)} = I(O_{ij} \leq c)\ ,\ c = 1, \ldots, C\ ,$$

so that $\mathbf{Y}_{ij} = (Y_{ij}^{(1)}, \ldots, Y_{ij}^{(C)})'$. Ordinal data is often modeled by assuming proportional odds (McCullagh, 1980; Stokes et al., 1995). For covariate vector $\mathbf{X}_{ij}$, the proportional odds model assumes that the $\mathbf{X}_{ij}$ effect is the same across levels of $O_{ij}$. This model for data $O_{ij}$ is specified by

$$\text{logit}\left( E[Y_{ij}^{(c)}|\mathbf{X}_{ij}] \right) \ = \ \delta_c + \mathbf{X}_{ij}'\boldsymbol{\beta}\ ,\ 1 \leq c \leq C,\ 1 \leq j \leq n_i\ . \tag{4.1}$$

The proportional odds assumption is easily relaxed by assuming a more general form for $E[Y_{ij}^{(c)}|\mathbf{X}_{ij}]$. Unless otherwise noted, (4.1) is assumed for the mean of $O_{ij}$. Interest here is in a marginal model for $O_{ij}$, so that only mean and variance structures are explicitly specified for $O_{ij}$ as in the generalized estimating equations of Liang and Zeger (1986).

Let $\boldsymbol{\mu}_{ij} = E[\mathbf{Y}_{ij}|\mathbf{X}_{ij}]$ and $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}', \ldots, \boldsymbol{\mu}_{in_i}')'$. For the observed vector $\mathbf{Y}_i = (\mathbf{Y}_{i1}', \ldots, \mathbf{Y}_{in_i}')'$, generalized estimating equations (Liang and Zeger, 1986) estimates the mean parameter $\boldsymbol{\beta}$ with the solution to

$$\mathbf{U}_{\boldsymbol{\beta}} = \sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}\ , \tag{4.2}$$

where $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$ and $\boldsymbol{\mu}_i$ is determined by (4.1). In addition to (4.1), another assumption determines the structure of variance matrix $\mathbf{V}_i \approx \text{var}(\mathbf{Y}_i)$. In alternating logistic regressions (ALR) for binary data, Carey et al. (1993) parameterized $\mathbf{V}_i$ in terms of odds ratios for response pairs. Heagerty and Zeger (1996) defined ALR for ordinal data by using odds ratios

$$\psi_{ijk}^{(a,b)} = \frac{P(Y_{ij}^{(a)}=1, Y_{ik}^{(b)}=1)\, P(Y_{ij}^{(a)}=0, Y_{ik}^{(b)}=0)}{P(Y_{ij}^{(a)}=1, Y_{ik}^{(b)}=0)\, P(Y_{ij}^{(a)}=0, Y_{ik}^{(b)}=1)},$$

for $1 \le a, b \le C$ and $1 \le j < k \le n_i$. ALR assumes a model for $\psi_{ijk}^{(a,b)}$ governed by a $q$ dimensional parameter $\boldsymbol{\alpha}$ for covariate vector $\mathbf{Z}_{ijk}^{(a,b)}$, where

$$\log\left(\psi_{ijk}^{(a,b)}\right) = \mathbf{Z}_{ijk}^{(a,b)\prime}\,\boldsymbol{\alpha}\,, \quad 1 \le j < k \le n_i\,, \quad 1 \le a, b \le C\,. \tag{4.3}$$

The model for data $O_{ij}$ is jointly specified by (4.1) and (4.3). The odds ratio $\psi_{ijk}^{(a,b)}$ determines the conditional expected value $\zeta_{ijk}^{(a,b)} = E[Y_{ij}^{(a)}|Y_{ik}^{(b)}]$ (Mardia, 1967). While estimating $\boldsymbol{\beta}$ through (4.2), ALR uses another set of estimating equations based on the conditional residuals $Y_{ij}^{(a)} - \zeta_{ijk}^{(a,b)}$ to estimate association parameter $\boldsymbol{\alpha}$. Let $\boldsymbol{\zeta}_{ijk}$ and $\boldsymbol{\zeta}_i$ be vectors of conditional expectations such that

$$\boldsymbol{\zeta}_{ijk} = (\zeta_{ijk}^{(1,1)}, \ldots, \zeta_{ijk}^{(1,C)}, \zeta_{ijk}^{(2,1)}, \ldots, \zeta_{ijk}^{(C,C)})'\,,$$

and

$$\boldsymbol{\zeta}_i = (\boldsymbol{\zeta}_{i12}', \ldots, \boldsymbol{\zeta}_{i1n_i}', \boldsymbol{\zeta}_{i23}', \ldots, \boldsymbol{\zeta}_{i(n_i-1)n_i}')'\,.$$

Also let the vector $\mathbf{Y}_i^*$ represent the observations associated with $\boldsymbol{\zeta}_i$, so that $\mathbf{Y}_{ij}^* = (\mathbf{Y}_{ij} \otimes \mathbf{1}_C)$ and $\mathbf{Y}_i^* = (\mathbf{Y}_{i1}^{*\prime}, \ldots, \mathbf{Y}_{i1}^{*\prime}, \mathbf{Y}_{i2}^{*\prime}, \ldots, \mathbf{Y}_{i(n_i-1)}^{*\prime})'$. As defined by Heagerty and

Zeger (1996) in ALR for multilevel data, $\boldsymbol{\alpha}$ is estimated by the solution to

$$\mathbf{U}_{\boldsymbol{\alpha},ALR} \;=\; \sum_{i=1}^{K} \partial\boldsymbol{\zeta}_i'/\partial\boldsymbol{\alpha}\, \mathrm{Diag}\{\boldsymbol{\zeta}_i(1-\boldsymbol{\zeta}_i)\}\; (\mathbf{Y}_i^* - \boldsymbol{\zeta}_i) \;=\; \mathbf{0}\,. \tag{4.4}$$

Estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are obtained by iterating between (4.2) and (4.4). The resulting estimates $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are consistent for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that for $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$, $\sqrt{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically multivariate normal with mean zero, given that (4.3) and (4.1) hold. If (4.3) does not hold, the outcome covariance is misspecified, and $\hat{\boldsymbol{\beta}}$ is still consistent for $\boldsymbol{\beta}$. Sandwich estimators based on (4.2) and (4.4) are available for the variances of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ (Heagerty and Zeger, 1996).

A generalization of ALR for binary data was defined by Zink and Qaqish (2009) resolving the dependence of variance estimates on observation order. While other marginal methods are available for analyzing correlated ordinal data (Liang, Zeger, and Qaqish, 1992; Molenberghs and Lesaffre, 1994), we will only consider the method defined by Zink and Qaqish (2009), called orthogonalized residuals. For a comprehensive review of marginal methods for correlated data, see Agresti (1999).

## 4.3 Orthogonalized residuals

Like ALR, orthogonalized residuals (ORTH) is an extension of Liang and Zeger's (1986) method, estimating marginal mean parameter $\boldsymbol{\beta}$ with the solution to (4.2). Unlike ALR, where association parameter $\boldsymbol{\alpha}$ estimation is based on conditional expectations $E[Y_{ij}^{(a)} | \, Y_{ik}^{(b)}]$, $\boldsymbol{\alpha}$ estimation is instead based on expectations of cross-products $Y_{ij}^{(a)} Y_{ik}^{(b)}$ conditional on $Y_{ij}^{(a)}$ and $Y_{ik}^{(b)}$, for $1 \leq a, b \leq C$ and $1 \leq j < k < n_i$. Define

$$\mu_{ijk}^{(a,b)} = E[\, Y_{ij}^{(a)}\, Y_{ik}^{(b)}\,]$$

and $\mu_{ij}^{(a)} = E[Y_{ij}^{(a)}]$. Let $T_{ijk}^{(a,b)}$ be a residual based on the conditional expectation $E[\, Y_{ij}^{(a)}\, Y_{ik}^{(b)} \mid Y_{ij}^{(a)},\, Y_{ik}^{(b)}\,]$, such that

$$\begin{aligned} T_{ijk}^{(a,b)} &= Y_{ij}^{(a)} Y_{ik}^{(b)} - E[\, Y_{ij}^{(a)}\, Y_{ik}^{(b)} \mid Y_{ij}^{(a)},\, Y_{ik}^{(b)}\,] \\ &= Y_{ij}^{(a)} Y_{ik}^{(b)} - \{\, \mu_{ijk}^{(a,b)} + b_{ijk:j}^{(a,b)}(Y_{ij}^{(a)} - \mu_{ij}^{(a)}) + b_{ijk:k}^{(a,b)}(Y_{ik}^{(b)} - \mu_{ik}^{(b)})\,\}\,, \end{aligned}$$

for

$$\begin{aligned} d_{ijk}^{(a,b)} &= \sigma_{ijj}^{(a)} \sigma_{ikk}^{(b)} - \sigma_{ijk}^{(a,b)\,2} \\ b_{ijk:j}^{(a,b)} &= \mu_{ijk}^{(a,b)}(1 - \mu_{ik}^{(b)})(\mu_{ik}^{(b)} - \mu_{ijk}^{(a,b)})/d_{ijk}^{(a,b)} \\ b_{ijk:k}^{(a,b)} &= \mu_{ijk}^{(a,b)}(1 - \mu_{ij}^{(a)})(\mu_{ij}^{(a)} - \mu_{ijk}^{(a,b)})/d_{ijk}^{(a,b)}\,, \end{aligned}$$

and

$$\sigma_{ijj}^{(a)} := \mathrm{var}(Y_{ij}^{(a)}) = \mu_{ij}^{(a)}(1 - \mu_{ij}^{(a)}) \qquad \sigma_{ijk}^{(a,b)} := \mathrm{cov}(Y_{ij}^{(a)}, Y_{ik}^{(b)}) = \mu_{ijk}^{(a,b)} - \mu_{ij}^{(a)} \mu_{ik}^{(b)}\,.$$

Let the vector $\mathbf{T}_i$ have elements $T_{ijk}^{(a,b)}$ such that

$$\mathbf{T}_i = (T_{i12}^{(1,1)}, \ldots, T_{i12}^{(1,C)}, T_{i12}^{(2,1)}, \ldots, T_{i(n_i-1)n_i}^{(C,C)})'\,.$$

For matrices $\mathbf{S}_i = E\,[-\partial \mathbf{T}_i / \partial \boldsymbol{\alpha}']$ and $\mathbf{P}_i \approx \mathrm{var}(\mathbf{T}_i)$, the ORTH estimate of $\boldsymbol{\alpha}$ is the solution to

$$\mathbf{U}_{\boldsymbol{\alpha},ORTH} = \sum_{i=1}^{K} \mathbf{S}_i' \mathbf{P}_i^{-1} \mathbf{T}_i = \mathbf{0}\,. \tag{4.5}$$

The variance of $\mathbf{T}_i$ is approximated by $\mathbf{P}_i$, having elements $\mathrm{cov}(\, T_{ijk}^{(a,b)}\,,\, T_{ij'k'}^{(c,d)}\,) = 0$ for $j \neq j'$ or $k \neq k'$, so that $\mathbf{P}_i$ is block diagonal. Here $a \wedge c = \min(a, c)$, and $\mathbf{P}_i$ has

nonzero elements $\text{cov}(\, T_{ijk}^{(a,b)} \,,\, T_{ijk}^{(c,d)} \,)$, $1 \leq a, b \leq C$ and $1 \leq c, d \leq C$, such that

$$\begin{aligned}
\text{cov}(\, T_{ijk}^{(a,b)} \,,\, T_{ijk}^{(c,d)} \,) \;=\; & \mu_{ijk}^{(a\wedge c,\, b\wedge d)} - b_{ijk:j}^{(a,b)}\mu_{ijk}^{(a\wedge c,\, d)} - b_{ijk:j}^{(c,d)}\mu_{ijk}^{(a\wedge c,\, b)} - b_{ijk:k}^{(a,b)}\mu_{ijk}^{(c,\, b\wedge d)} \\
& -b_{ijk:k}^{(c,d)}\mu_{ijk}^{(a,\, b\wedge d)} + (\mu_{ij}^{(c)} b_{ijk:j}^{(c,d)} + \mu_{ik}^{(d)} b_{ijk:k}^{(c,d)} - \mu_{ijk}^{(c,d)})\mu_{ijk}^{(a,b)} \\
& +(\mu_{ij}^{(a)} b_{ijk:j}^{(a,b)} + \mu_{ik}^{(b)} b_{ijk:k}^{(a,b)} - \mu_{ijk}^{(a,b)})\mu_{ijk}^{(c,d)} + b_{ijk:j}^{(a,b)} b_{ijk:k}^{(c,d)} \mu_{ijk}^{(a,d)} \\
& +b_{ijk:j}^{(c,d)} b_{ijk:k}^{(a,b)} \mu_{ijk}^{(c,b)} + b_{ijk:j}^{(a,b)} b_{ijk:j}^{(c,d)} \mu_{ij}^{(a\wedge c)} + b_{ijk:k}^{(a,b)} b_{ijk:k}^{(c,d)} \mu_{ik}^{(b\wedge d)} \\
& -(\mu_{ij}^{(a)} b_{ijk:j}^{(a,b)} + \mu_{ik}^{(b)} b_{ijk:k}^{(a,b)})(\mu_{ij}^{(c)} b_{ijk:j}^{(c,d)} + \mu_{ik}^{(d)} b_{ijk:k}^{(c,d)}) + \mu_{ijk}^{(a,b)} \mu_{ijk}^{(c,d)} \,.
\end{aligned} \tag{4.6}$$

For this approximation of $\text{var}(\mathbf{T}_i)$, ORTH estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are obtained by iterating between (4.2) and (4.5). The resulting $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are consistent for $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$, such that $\sqrt{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically multivariate normal with mean zero, given that the model defined by (4.3) and (4.1) holds. A heuristic argument for the asymptotic distribution of $\sqrt{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is given in an appendix. Also addressed in an appendix is the derivation of (4.6), and (4.6) in the case of binary responses ($C = 1$), which reduces to the ORTH defined by Zink and Qaqish (2009).

For multilevel data ($C > 1$), it can be shown that when $\mathbf{P}_i$ is a diagonal matrix with non-zero elements ($a = c$, $b = d$) given by (4.6), the resulting $\hat{\boldsymbol{\alpha}}$ is the same as that in ALR for multilevel data defined by Heagerty and Zeger (1996). The block diagonal matrices in $\mathbf{P}_i$ have off diagonal elements defined in (4.6) that account for the variance between $T_{ijk}^{(a,b)}$ and $T_{ijk}^{(c,d)}$, when $a \neq c$ or $b \neq d$. These are cross product residuals concerning the same pair of multinomial observations, $O_{ij}$ and $O_{ik}$. Because $Y_{ij}^{(a)}$ and $Y_{ij}^{(c)}$ are correlated, as well as $Y_{ik}^{(b)}$ and $Y_{ik}^{(d)}$, the actual covariance between $T_{ijk}^{(a,b)}$ and $T_{ijk}^{(c,d)}$ is nonzero. Therefore the working covariance in $\mathbf{U}_{\boldsymbol{\alpha},\, ORTH}$ is closer to the actual covariance than that in $\mathbf{U}_{\boldsymbol{\alpha},\, ALR}$, meaning that the ORTH $\hat{\boldsymbol{\alpha}}$ should be more efficient than the ALR $\hat{\boldsymbol{\alpha}}$ by the tenets of estimating function theory (Qin and Lawless, 1994).

Both orthogonalized residuals and alternating logistic regressions have inefficient

estimates of $\boldsymbol{\alpha}$ relative to second order estimating equations (Liang et al., 1992; Zink and Qaqish, 2009), which estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ parameters simultaneously. There is a considerable computational advantage to estimating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ separately, where matrices of order $n^2$ are inverted, instead of inverting matrices of order $n^4$ as in second order estimating equations, for clusters of size $n$ (Carey et al., 1993).

Although it is expected that ORTH and ALR would be less efficient than second order estimating equations, it is also expected that ORTH could gain efficiency in $\hat{\boldsymbol{\alpha}}$ compared to ALR, by assuming a non-diagonal structure for the covariance of $\mathbf{T}_i$ in (4.5), and also because the residual $\mathbf{T}_i$ has a small correlation with $\mathbf{Y}_i$ (Zink and Qaqish, 2009). In addition to this increased precision, the method proposed here has several advantages relative to the method proposed by Heagerty and Zeger (1996). First, unlike the Heagerty and Zeger (1996) estimate, the resulting $\widehat{\text{var}}(\hat{\boldsymbol{\alpha}})$ based on (4.5) is invariant to the ordering of elements within cluster. The ORTH $\hat{\boldsymbol{\alpha}}$ is also estimated with a non-stochastic derivative matrix, allowing the application of standard estimating equation theory.

## 4.4 Example

The methods described above will be illustrated in an analysis of clinical trial data from the sensory retraining study. This study was designed to compare the perceived altered sensation for post operative patients in two treatment groups. Both groups received the standard treatment following a bilateral sagittal split osteotomy, a surgical procedure on the mandible, while one group also received the treatment of sensory retraining exercises (Phillips et al., 2007). Ordered outcomes for altered sensation were measured at 6, 13, and 26 weeks after surgery. There were 93 patients in the standard treatment group and 91 patients in the sensory retraining group. Overall 178 (97%) patients had observed outcomes at all visits.

The following model represents multiple categorical outcomes observed over time, where $j$ indexes outcome and $t$ indexes observation time, with response $O_{ijt} = c$ for some $c \in 1, \ldots, C+1$. Let $Y_{ijt}^{(c)}$ represent $O_{ijt}$ as binary element

$$Y_{ijt}^{(c)} = I(O_{ijt} \leq c)\ ,\ c = 1, \ldots, C\ .$$

For the example of sensory retraining data, $O_{ijt}$ is an ordinal measurement of altered sensation with seven levels ($C = 6$) where $O_{ijt} = 1$ indicated the most favorable outcome and $O_{ijt} = 7$ the least favorable. These seven levels were collapsed into three levels ($C = 2$) for this analysis.

Perceived altered sensation was measured with $O_{ijt}$ where $O_{i1t}$ ($j = 1$) measured the loss of lip sensitivity, $O_{i2t}$ ($j = 2$) measured the level of unusual feelings, and $O_{i3t}$ ($j = 3$) measured numbness. Each of these responses was recorded before surgery and at three subsequent times. The initial responses are not used in this analysis because there is very little variation in pre-surgical measurements of altered sensation. Time or visit is indexed by $t$, for one ($t = 1$), three ($t = 2$), and six ($t = 3$) months after surgery. The following model definitions include covariate $T_i$, an indicator that subject $i$ received the experimental sensory retraining treatment.

In the analysis of the sensory retraining data, the marginal mean of binary indicators $Y_{ijt}^{(c)}$ are restricted by

$$\begin{aligned} \text{logit}\left( E[Y_{ijt}^{(c)} | \mathbf{X}_{ij}] \right) \ &= \ \delta_c + \beta_{0j} + \beta_{1t} + \beta_{2t} T_i + \beta_{1jt} + \beta_{2jt} T_i + \\ &\quad \beta_3 G_i + \beta_4 J_i\ , \end{aligned} \tag{4.7}$$

for $1 \leq c \leq C$ and $1 \leq j, t \leq 3$ with the restriction that $\beta_{01} = \beta_{11} = \beta_{11t} = \beta_{21t} = 0$. This model is saturated in time and treatment for each outcome.

Note that this is a proportional odds model, where the response/covariate relation-

ship is independent of response level $c$. Covariate $T_i = 1$ if subject $i$ was in the sensory retraining group, and 0 otherwise, $G_i = 1$ if subject $i$ received genioplasty as part of surgery (0 otherwise), and $J_i = 1$ if the surgery for subject $i$ involved only one jaw (0 otherwise) (Phillips et al., 2007).

This model allows that the marginal mean of each outcome vary by treatment group over time. This model also allows that the odds of altered sensation differ initially by genioplasty or number of jaws in surgery, however, the corresponding rate of change in the odds of altered sensation over time is not allowed to vary in this model.

In addition to (4.7), the model for $Y_{ijt}^{(c)}$ includes a restriction on the pairwise odds ratio between $Y_{ijt}^{(a)}$ and $Y_{iks}^{(b)}$, where indices $j, k$ represent outcome, $s, t$ index observation time, and $1 \leq a, b \leq C$. A preliminary model saturated by outcome pair $(j, k)$ can be written

$$\begin{aligned} \log\left(\psi_{ijk,st}^{(a,b)}\right) &= \alpha_{012} I(j=1,k=2;s=t) + \alpha_{013} I(j=1,k=3;s=t) + \\ &\quad \alpha_{023} I(j=2,k=3;s=t) + \alpha_{111} I(j=1,k=1;s\neq t) + \\ &\quad \alpha_{122} I(j=2,k=2;s\neq t) + \alpha_{133} I(j=3,k=3;s\neq t) + \\ &\quad \alpha_{112} I(j=1,k=2;s\neq t) + \alpha_{113} I(j=1,k=3;s\neq t) + \\ &\quad \alpha_{123} I(j=2,k=3;s\neq t)\,, \end{aligned} \tag{4.8}$$

for $j \leq k$ and $s \leq t$, omitting the case that $j = k$ and $s = t$. In a Wald test that $\alpha_{112} = \alpha_{113} = \alpha_{123}$ and $\alpha_{111} = \alpha_{122} = \alpha_{133}$, the observed test statistic is 2.50 ($p = 0.645$) with four degrees of freedom. This test result indicates that associations between outcomes observed at different times can be expressed with two odds ratios, one representing the same outcome observed at different times, and the other representing

different altered sensation outcomes observed at different times, reducing (4.8) to

$$\begin{aligned} \log\left(\psi_{ijk,st}^{(a,b)}\right) &= \alpha_{012} I(j=1,k=2;s=t) + \alpha_{013} I(j=1,k=3;s=t) + \\ &\quad \alpha_{023} I(j=2,k=3;s=t) + \alpha_1 I(j=k;s\neq t) + \alpha_2 I(j\neq k;s\neq t)\,, \end{aligned} \tag{4.9}$$

for $j \leq k$ and $s \leq t$, except the case that $j = k$ and $s = t$. The Wald test that $\alpha_{012} = \alpha_{013} = \alpha_{023}$ has test statistic 14.30 ($p < 0.001$) with two degrees of freedom, indicating that further model reduction is not appropriate.

The above model for the pairwise odds ratio $\psi_{ijk,st}^{(a,b)}$ allows the association between $Y_{ijt}^{(a)}$ and $Y_{iks}^{(b)}$ to vary by response pair $(j,k)$ within time, and specifies two log odds ratios for response pairs between times. In the model with restrictions (4.7) and (4.9), to test whether the effect of time and treatment vary for different outcomes, a Wald test that $\beta_{1jt} = 0$ for $j,t = 2,3$ and that $\beta_{2jt} = 0$ for $j = 2,3$, $1 \leq t \leq 3$ can be used. This test has observed statistic 7.87 with ten degrees of freedom and $p = 0.642$, so the marginal mean model can be reduced to

$$\begin{aligned} \text{logit}\left(E[Y_{ijt}^{(c)}|\mathbf{X}_{ij}]\right) &= \delta_c + \beta_{0j} + \beta_{1t} + \beta_{2t} T_i + \\ &\quad \beta_3 G_i + \beta_4 J_i\,, \end{aligned} \tag{4.10}$$

for $1 \leq c \leq C$ and $1 \leq j,t \leq 3$ with the restriction that $\beta_{01} = \beta_{11} = 0$. The resulting parameter estimates for this model using both ALR as defined by Heagerty and Zeger (1996) and the proposed orthogonalized residuals for ordinal data are in Table 4.1.

Note that positive parameter estimates from (4.10) in Table 4.1 are associated with smaller, or more favorable, values of $O_{ijt}$. Note also that although the change over time of different altered sensation measurements did not vary, the measurements themselves varied, with patients most likely to have favorable (small) measurements for unusual feelings ($\beta_{02}$), and least likely to have favorable measurements for numbness ($\beta_{03}$).

The effect of the sensory retraining exercises on altered sensation relative to standard treatment is represented by $\beta_{2t}$, $t = 1, 2, 3$. Although none of these parameter estimates are significantly different from zero by themselves (at 0.05), a Wald test that $\beta_{2t} = 0$ for $t = 1, 2, 3$ has an observed statistic of 10.67 with three degrees of freedom and $p = 0.014$. This result indicates that the course of altered sensation over time for the two treatment groups was significantly different.

The parameter estimates for the odds ratio model are also in Table 4.1. Given the log scale, these estimates represent odds ratios from 6.5 ($\exp\{\hat{\alpha}_{012}\}$) to 15.2 ($\exp\{\hat{\alpha}_{023}\}$). The parameter $\alpha_{012}$ is the log odds of having more lip sensitivity (i.e. a more favorable result) given that unusual feelings are reduced within observation time, while $\alpha_{023}$ is the log odds of having reduced unusual feelings given that numbness is reduced within observation. Likewise, $\alpha_{013}$ is the log odds of increased lip sensitivity given that numbness is reduced within visit, and $\alpha_1$ is the log odds that one altered sensation measurement will be lower given that another is lower at a different visit. The orthogonalized residuals estimate of $\boldsymbol{\alpha} = (\alpha_{012},\ \alpha_{013},\ \alpha_{023},\ \alpha_1,\ \alpha_2)'$ corresponds to the odds ratios (95% confidence intervals): 6.52 (4.37,9.72), 15.69 (9.88,24.92), 12.27 (7.70,19.56), 6.13 (4.51,8.33), and 3.87 (2.90,5.15) for $\hat{\alpha}_{012}$, $\hat{\alpha}_{013}$, $\hat{\alpha}_{023}$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

Also of interest in this analysis is whether the precision of association parameter estimates in the orthogonalized residual formulation is improved relative to ALR as proposed by Heagerty and Zeger (1996). The ratios of estimated variances for parameter estimates in Table 4.1 are 101, 85, 87, 93 and 91% for $\hat{\alpha}_{012}, \hat{\alpha}_{013}, \hat{\alpha}_{023}$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$, respectively, where a ratio of less than 100% indicates that the variance estimated with the orthogonalized residuals is smaller than that for ALR.

## 4.5 Simulation

A small simulation experiment was conducted to investigate the relative efficiency of ALR to ORTH in a finite sample modeled similarly to the sensory retraining study. Categorical response data $O_{ij}$ will be generated with three ($C = 2$) levels for subject $i = 1, \ldots, K$ and outcome $j = 1, 2, 3$. Generated data will be similar to that observed in the sensory retraining study, with the marginal mean of $Y_{ij}^{(c)} = I(O_{ij} \leq c)$ dependent on $j$ and dichotomous covariate $T_i$, indicating treatment group, such that

$$\text{logit}\left( E[Y_{ij}^{(c)}|\mathbf{X}_{ij}] \right) = \delta_c + \beta_{0j} + \beta_{2j} T_i \,, \tag{4.11}$$

for $j = 1, 2, 3$ and $c = 1, 2$, where $\beta_{01} = 0$. Half of all subjects will have $T_i = 1$, and the remaining half will have $T_i = 0$. This data generating model represents the full marginal mean (4.7) considered in the sensory retraining analysis at one observation time, disregarding covariates other than $T_i$. For $\beta_{2j} = 0$, $j = 1, 2, 3$, there is no effect of $T_i$. For $\beta_{21} = \beta_{22} = \beta_{23}$, the effect of $T_i$ is the same across outcome $j$.

In addition to the marginal mean (4.11), the data generating mechanism includes the association between outcomes $Y_{ij}^{(c)}$ and $Y_{ik}^{(d)}$ within subject, specified by the log odds ratio

$$\begin{aligned} \log\left(\psi_{ijk}^{(a,b)}\right) &= \alpha_{012} I(j = 1, k = 2) + \alpha_{013} I(j = 1, k = 3) + \\ &\quad \alpha_{023} I(j = 2, k = 3) \,, \end{aligned} \tag{4.12}$$

for $1 \leq j < k \leq 3$, $a = 1, 2$ and $b = 1, 2$. Data will be generated by specifying the first two moments of $(O_{i1}, O_{i2}, O_{i3})'$ with parameters $\boldsymbol{\delta} = (\delta_1, \delta_2)'$, $\boldsymbol{\beta} = (\beta_{02}, \beta_{03}, \beta_{21}, \beta_{22}, \beta_{23})'$, and $\boldsymbol{\alpha} = (\alpha_{012}, \alpha_{013}, \alpha_{023})'$ taken from an analysis of the sensory retraining data at the last observation time, where $\hat{\boldsymbol{\delta}} = (-0.5, 2.4)'$, $\hat{\boldsymbol{\beta}} = (0.5, -0.6, 0.6, 0.2, 0.4)'$

and $\hat{\boldsymbol{\alpha}} = (2.2, 2.9, 3.0)'$. Data generation is based on the method defined by Gange (1995) for correlated ordinal data given these $\boldsymbol{\delta}$, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$, for $K = 100$ or 200 subjects. The correct models (4.11) and (4.12) are used to analyze each of one thousand realizations, and $\boldsymbol{\alpha}$ and the standard errors of $\hat{\boldsymbol{\alpha}}$ will be estimated with both orthogonalized residuals and ALR for ordinal data as defined by Heagerty and Zeger (1996).

In order to evaluate the relative performance of orthogonalized residuals and ALR, they will be compared in the estimated bias of $\hat{\boldsymbol{\alpha}}$, as well as the Monte Carlo standard errors of $\hat{\boldsymbol{\alpha}}$. The relative efficiency of $\hat{\boldsymbol{\alpha}}$ from the two methods in particular is of interest, and will be approximated here by the ratio of Monte Carlo variances of $\hat{\boldsymbol{\alpha}}$, where the Monte Carlo variance of $\alpha_{0jk}$ is

$$\sum_{s=1}^{1000} (\hat{\alpha}_{0jk}^{(s)} - \bar{\alpha}_{0jk})^2 / 999 \,,$$

for $\bar{\alpha}_{0jk} = \sum_{s=1}^{1000} \hat{\alpha}_{0jk}^{(s)} / 1000$. In addition to these results for $\hat{\boldsymbol{\alpha}}$, results relating to the estimated standard errors of $\hat{\boldsymbol{\alpha}}$ will also be included. The bias of the standard error estimates relative to the Monte Carlo standard errors of $\hat{\boldsymbol{\alpha}}$ will be estimated, as well as the coverage of nominally 95% confidence intervals for $\hat{\boldsymbol{\alpha}}$, estimated by the percent of 95% confidence intervals that include the true value of $\boldsymbol{\alpha}$.

### 4.5.1 Simulation results

The bias and Monte Carlo standard error of $\hat{\boldsymbol{\alpha}}$ is in Table 4.2. Note that the bias of $\hat{\boldsymbol{\alpha}}$ is reduced with increasing $K$ (as expected), and that the bias of ORTH is always less than the bias of ALR, although the percent relative bias of $\hat{\boldsymbol{\alpha}}$ never exceeds 4% for either method. The Monte Carlo standard error for $\hat{\boldsymbol{\alpha}}$ in ORTH is also always less than that for ALR. The estimated efficiency of ALR relative to ORTH ranges from 92 to 97%.

The bias of standard error estimators and the coverage of nominally 95% confidence intervals are shown in Table 4.3. Although the difference in coverage rates of ORTH and ALR confidence intervals is not large, it is notable that the coverage for ORTH is always the same or larger than for ALR. Also note that the bias of standard error estimators is always less than zero, so that the standard error of $\hat{\boldsymbol{\alpha}}$ was underestimated by both ORTH and ALR. The bias of the ORTH standard error estimator, however, is always closer to zero than the bias of the ALR estimator. With respect to the Monte Carlo standard errors, the percents relative bias of the ALR standard error estimates are -15, -11, -10, -6, -6, and -14%, while those for the ORTH standard error estimates are -13, -10, -7, -4, -5, and -12%.

## 4.6 Conclusions

Data was simulated so that the circumstances of analysis would resemble the sensory retraining clinical trial at the last observation time, where cluster sizes are small and associations within clusters are large. The simulated data may reasonably characterize the sensory retraining data but was not intended to characterize a broad range of analyses with ordinal data. Data with small associations between clustered observations, for instance, were not considered, and neither was data with large clusters or with varying cluster sizes.

Large cluster sizes in particular can affect analysis results and can pose a significant challenge in accounting for associations. A small efficiency gain was noted here for simulated data even with a cluster size of three, and an estimated gain in precision in the analysis of the sensory retraining data suggest that more efficiency might be gained with larger cluster sizes for ORTH relative to ALR as defined by Heagerty and Zeger (1996). In efficiency and all other measures considered here, the association parameter estimated by ORTH was shown to have at least as desirable and often more desirable

Table 4.1: Estimates and estimated standard errors for marginal mean and pairwise log odds ratio parameters with ALR and orthogonalized residuals in the Sensory Retraining example. ALR estimates were calculated by the Heagerty and Zeger (1996) formulation, and the ORTH estimates are those proposed above.

| | ALR | | ORTH | |
|---|---|---|---|---|
| $\delta_1$ | -1.956 | 0.2394 | -1.951 | 0.2392 |
| $\delta_2$ | 0.633 | 0.2294 | 0.635 | 0.2293 |
| $\beta_{02}$ | 0.234 | 0.0999 | 0.233 | 0.1000 |
| $\beta_{03}$ | -0.795 | 0.0947 | -0.795 | 0.0947 |
| $\beta_{12}$ | 0.667 | 0.1299 | 0.661 | 0.1296 |
| $\beta_{13}$ | 1.392 | 0.1749 | 1.387 | 0.1748 |
| $\beta_{21}$ | -0.302 | 0.2394 | -0.303 | 0.2394 |
| $\beta_{22}$ | 0.369 | 0.2425 | 0.370 | 0.2426 |
| $\beta_{23}$ | 0.362 | 0.2388 | 0.362 | 0.2389 |
| $\beta_3$ | 0.324 | 0.2099 | 0.324 | 0.2100 |
| $\beta_4$ | 0.433 | 0.2025 | 0.434 | 0.2025 |
| $\alpha_{012}$ | 1.927 | 0.2035 | 1.874 | 0.2043 |
| $\alpha_{013}$ | 2.399 | 0.2580 | 2.508 | 0.2378 |
| $\alpha_{023}$ | 2.762 | 0.2523 | 2.753 | 0.2359 |
| $\alpha_1$ | 1.740 | 0.1623 | 1.814 | 0.1565 |
| $\alpha_2$ | 1.376 | 0.1532 | 1.353 | 0.1464 |

qualities than the ALR estimate, including bias in $\hat{\boldsymbol{\alpha}}$ and the standard error estimator, and coverage of 95% confidence intervals.

An advantage of ORTH over ALR is that the representation of the estimating equations for $\boldsymbol{\alpha}$ in a standard form (i.e., equation (4.5)) permitted the use of non-diagonal working covariance matrix $\mathbf{P}_i$ to provide a more efficient estimator of $\boldsymbol{\alpha}$. We conjecture that further efficiency gains are possible by introducing non-zero off-diagonal block elements of $\mathbf{P}_i$, for example, through a working correlation structure as proposed by Zink and Qaqish (2009) for binary data.

Table 4.2: ALR and ORTH estimates of association parameter $\boldsymbol{\alpha} = (\alpha_{012}, \alpha_{013}, \alpha_{023})'$ for simulated ordinal data. For data with $K = 100$ or 200 clusters, each with $j = 3$ outcomes, having marginal mean (4.11) and association (4.12).

| | Bias of $\hat{\boldsymbol{\alpha}}$ | | Monte Carlo SE | | ALR Est. |
|---|---|---|---|---|---|
| $K$ | ALR | ORTH | ALR | ORTH | Efficiency |
| | | | $\hat{\alpha}_{012}$ | | |
| 100 | 0.038 | 0.027 | 0.519 | 0.508 | 95.8 |
| 200 | 0.019 | 0.015 | 0.353 | 0.347 | 96.6 |
| | | | $\hat{\alpha}_{013}$ | | |
| 100 | 0.077 | 0.067 | 0.581 | 0.564 | 94.2 |
| 200 | 0.037 | 0.031 | 0.389 | 0.374 | 92.4 |
| | | | $\hat{\alpha}_{023}$ | | |
| 100 | 0.108 | 0.098 | 0.633 | 0.623 | 96.9 |
| 200 | 0.063 | 0.054 | 0.480 | 0.465 | 93.8 |

Table 4.3: Coverage of ALR and ORTH confidence intervals for $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_{012}, \hat{\alpha}_{013}, \hat{\alpha}_{023})'$, as well as the average bias of the associated standard error estimates, in simulated ordinal data. For data with $K = 100$ or 200 clusters, each with $j = 3$ outcomes, having marginal mean (4.11) and association (4.12).

| | Coverage of 95% CI | | Bias of $\widehat{\text{SE}}$ | |
|---|---|---|---|---|
| $K$ | ALR | ORTH | ALR | ORTH |
| | | $\hat{\alpha}_{012}$ | | |
| 100 | 91.1 | 92.2 | -0.0769 | -0.0646 |
| 200 | 92.9 | 92.9 | -0.0382 | -0.0351 |
| | | $\hat{\alpha}_{013}$ | | |
| 100 | 93.7 | 94.1 | -0.0556 | -0.0421 |
| 200 | 94.6 | 94.8 | -0.0238 | -0.0147 |
| | | $\hat{\alpha}_{023}$ | | |
| 100 | 95.2 | 95.5 | -0.0355 | -0.0289 |
| 200 | 92.6 | 93.3 | -0.0673 | -0.0569 |

# Summary and Future Research

## 5.1 Summary of research

### 5.1.1 Semi-parametric efficient estimation for incomplete longitudinal binary data

For our first research topic we defined a specific form for the asymptotically semi-parametric efficient estimator for longitudinal binary data. This estimator was applied to data from a fifteen year cohort survey, and the efficiency of weighted generalized estimating equations, a computationally simple inverse probability weighted estimator, was assessed relative to that estimator for longitudinal binary data with dropout. We show that there is efficiency to be gained upon generalized estimating equations and the weighted generalized estimating equations in the presence of incomplete data. Although the computation of the semi-parametric efficient estimator is not necessarily straightforward, especially for large clusters, the percent efficiency gain can be significant depending on the nature of the data being analyzed. For small clusters, in the case where dropout rate and correlation is high, efficiency can be markedly increased even under circumstances where generalized estimating equations is consistent for marginal mean parameters.

### 5.1.2 Alternating logistic regressions with improved finite sample properties

We proposed adjustments to alternating logistic regressions and illustrated their use with data from a cluster survey repeated over time, where cluster sizes are large and associations within clusters are small. Analysis for cluster survey data often includes an interest in association parameter inference, where alternating logistic regressions is especially useful. The proposed adjustment to the ALR estimating equations did not always provide association parameter estimates with less bias than standard ALR, however, the reduction in bias was often marked. In addition, the variance estimators in the adjusted estimating equations were less biased than those of standard ALR.

Variance estimates were additionally improved by using an extension of the bias-corrected variance estimator introduced by Mancl and DeRouen (2001). Results with simulated data indicate that the proposed small sample adjusted ALR is an appropriate analysis for the chosen subset of underage drinking data, and possibly for other repeated cluster survey data. Overall we have shown that ALR can be applied with increased accuracy to small samples, and that the proposed method improves inference for association parameters in repeated binary data when applying alternating logistic regressions.

### 5.1.3 Orthogonalized residuals for ordinal data

We proposed an extension of an alternate formulation of alternating logistic regressions for ordinal data. This alternate formulation of ALR improves upon that defined by Carey et al. (1993) by resolving the dependence of the sandwich variance estimate on observation order, and represents the association estimating equations in the standard estimating equations format. In addition to these improvements, the proposed method

was shown to increase the efficiency with which association parameters were estimated with simulated data relative to the ALR for ordinal data proposed by Heagerty and Zeger (1996). In efficiency and also for other considered measures, including coverage of approximate confidence intervals, the association parameters estimated by ORTH were shown to have at least as desirable and often more desirable qualities than the ALR estimate defined by Heagerty and Zeger (1996).

## 5.2 Future research

### 5.2.1 Semi-parametric efficient estimation for incomplete longitudinal binary data

A distinct disadvantage of the weighted GEE approaches we considered is their requirement that incomplete data be monotonically missing, and thus do not use all available data. There are existing methods for intermittently missing data, including multiple imputation (Paik, 1997), and those of Lin et al. (2004), who propose a class of inverse intensity-of-visit process-weighted estimators in marginal regression models that allow for arbitrary patterns of missing data. A possible extension of our proposed procedure would be to impute response data at the intermittently missing time points, prior to application of the inverse-probability weighted semi-parametric efficient estimator.

### 5.2.2 Alternating logistic regressions with improved finite sample properties

The variations considered in our analysis with alternating logistic regressions and simulated data may fairly represent a cluster survey repeated over time, but there are circumstances where alternating logistic regressions are often used that cannot be rep-

resented or interpolated from our results. Large associations in particular were not considered, and neither were data with small clusters or with varying cluster sizes. The proposed adjustments were shown in our simulation study to be useful in that they estimate association parameters with less bias than standard ALR, consistently across $K$ and for varying marginal means, but may not be the most desirable for small cluster sizes or when associations are relatively large. The most effective analysis of correlated binary data for making inference on association parameters in this setting is an outstanding issue. The usefulness and possible improvement upon standard ALR in this case would complement our research so far on related topics.

### 5.2.3 Orthogonalized residuals for ordinal data

We observed a small efficiency gain for simulated data for small cluster sizes, and an estimated gain in precision in the analysis of the sensory retraining data suggest that more efficiency might be gained with larger cluster sizes for ORTH relative to ALR as defined by Heagerty and Zeger (1996). Additional efficiency could be gained from a different specification of the covariance matrix in the association estimating equations. In orthogonalized residuals Zink and Qaqish (2009) for binary data, the variance in the association estimating equations was defined such that a non-zero exchangeable correlation could be estimated. This correlation would not necessarily be of analytic interest, although its estimation for ordinal data could improve further upon the efficiency of association parameter estimates by more accurately representing the actual variance of the association residual.

# Appendix

## Asymptotic distribution of ORTH estimators

The distribution of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$, is developed, in the proposed method of orthogonalized residuals for ordinal data. Marginal mean parameter $\boldsymbol{\beta}$ is a $p \times 1$ vector, and $\boldsymbol{\alpha}$ is a $q \times 1$ vector of marginal association parameters. The estimating equations for marginal mean and association parameters are

$$\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \sum_{i=1}^{K} \mathbf{U}_{\boldsymbol{\beta} i}(\boldsymbol{\theta}) = \sum_{i=1}^{K} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}\,,$$

$$\mathbf{U}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \sum_{i=1}^{K} \mathbf{U}_{\boldsymbol{\alpha} i}(\boldsymbol{\theta}) = \sum_{i=1}^{K} \mathbf{S}_i' \mathbf{P}_i^{-1} \mathbf{T}_i \;=\; \mathbf{0}\,.$$

For $\mathbf{U}_{\boldsymbol{\beta}}$, matrix $\mathbf{D}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ and $\mathbf{V}_i = \mathrm{var}(\mathbf{Y}_i)$. For estimating equation $\mathbf{U}_{\boldsymbol{\alpha}}$, $\mathbf{P}_i \approx \mathrm{var}(\mathbf{T}_i)$, and the vector $\mathbf{T}_i$ has elements $T_{ijk}^{(a,b)}$ such that

$$T_{ijk}^{(a,b)} = Y_{ij}^{(a)} Y_{ik}^{(b)} - \{\, \mu_{ijk}^{(a,b)} + b_{ijk:j}^{(a,b)} (Y_{ij}^{(a)} - \mu_{ij}^{(a)}) + b_{ijk:k}^{(a,b)} (Y_{ik}^{(b)} - \mu_{ik}^{(b)}) \,\}\,.$$

The partial derivative matrix $\mathbf{S}_i = E\left[-\partial \mathbf{T}_i / \partial \boldsymbol{\alpha}'\right] =$ is defined by vectors $E\left[-\partial T_{ijk}^{(a,b)} / \partial \boldsymbol{\alpha}\right]$ such that

$$E\left[-\frac{\partial T_{ijk}^{(a,b)}}{\partial \boldsymbol{\alpha}}\right] \;=\; \left\{ \frac{1}{\mu_{ijk}^{(a,b)}} + \frac{1}{\mu_{ij}^{(a)} - \mu_{ijk}^{(a,b)}} + \frac{1}{\mu_{ik}^{(b)} - \mu_{ijk}^{(a,b)}} + \frac{1}{1 - \mu_{ij}^{(a)} - \mu_{ik}^{(b)} + \mu_{ijk}^{(a,b)}} \right\}^{-1} \mathbf{Z}_{ijk}'\,.$$

The following is a general heuristic argument that $\sqrt{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically multivariate normal. For a more technical illustration, see pages 76-78 of the dissertation of Richard Zink (2003) for the asymptotic behavior of orthogonalized residuals for binary

data. Define $\mathbf{U}_i(\boldsymbol{\theta}) = (\mathbf{U}'_{\boldsymbol{\beta}i}(\boldsymbol{\theta}), \mathbf{U}'_{\boldsymbol{\alpha}i}(\boldsymbol{\theta}))'$ and $\mathbf{U}(\boldsymbol{\theta}) = \sum_{i=1}^{K} \mathbf{U}_i(\boldsymbol{\theta})$.

*Theorem 1.* $\sqrt{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to $MVN\,(\mathbf{0}, \Sigma)$ for matrix $\Sigma$, as $K \to \infty$.

*Corollaries:*

*(i)* $\sqrt{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\text{p}} -\sqrt{K}\left[\frac{\partial}{\partial \boldsymbol{\theta}}\,\mathbf{U}(\boldsymbol{\theta})\,\right]^{-1}\mathbf{U}(\boldsymbol{\theta})$.

*(ii)* $K^{-\frac{1}{2}}\mathbf{U}(\boldsymbol{\theta})$ converges in distribution to MVN $\left(\mathbf{0}, \lim_{K\to\infty} K^{-1}\sum_{i=1}^{K}\Gamma_i\right)$, for $\Gamma_i = \text{var}(\,\mathbf{U}_i(\boldsymbol{\theta})\,)$.

*(iii)* $K^{-1}\,\frac{d}{d\boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta})$ converges in probability to $K^{-1}\,D$ , where $D = E\left[\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta})\right]$.

*Proof Theorem 1 .* Note that

$$\begin{aligned}
\sqrt{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \;&\xrightarrow{\text{p}}\; -\sqrt{K}\left[\frac{\partial}{\partial \boldsymbol{\theta}}\,\mathbf{U}(\boldsymbol{\theta})\,\right]^{-1}\mathbf{U}(\boldsymbol{\theta}) \\
&=\; -\sqrt{K}\left[K^{-1}\frac{\partial}{\partial \boldsymbol{\theta}}\,\mathbf{U}(\boldsymbol{\theta})\,\right]^{-1}K^{-1}\mathbf{U}(\boldsymbol{\theta}) \\
&=\; -\left[K^{-1}\frac{\partial}{\partial \boldsymbol{\theta}}\,\mathbf{U}(\boldsymbol{\theta})\,\right]^{-1}K^{-\frac{1}{2}}\mathbf{U}(\boldsymbol{\theta})\,. \qquad \text{(A.1)}
\end{aligned}$$

Given *(ii)* and *(iii)* and by Slutsky's Theorem (Sen and Singer (1993), Theorem 3.4.3), (A.1) converges in distribution to MVN $\left(\mathbf{0},\, \lim_{K\to\infty} KD^{-1}\left(\sum_{i=1}^{K}\Gamma_i\right)D^{-1}\right)$, so

$$\sqrt{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \text{MVN}\left(\mathbf{0},\, \lim_{K\to\infty} KD^{-1}\left(\sum_{i=1}^{K}\Gamma_i\right)D^{-1}\right)\,.$$

□

*Corollary (i).* Expanding $\mathbf{U}(\hat{\boldsymbol{\theta}})$ about $\hat{\boldsymbol{\theta}}$,

$$\mathbf{0} \;=\; \mathbf{U}(\hat{\boldsymbol{\theta}}) \;\approx\; \mathbf{U}(\boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}}\,\mathbf{U}(\boldsymbol{\theta})\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \frac{1}{2}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\left[\frac{\partial^2}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}\,\mathbf{U}(\boldsymbol{\theta})\right](\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\,.$$

Under regularity conditions, the third term on the right converges in probability to $\mathbf{0}$ by the Weak Law of Large Numbers (WLLN; Sen and Singer (1993), Theorem 2.3.7), so that

$$\begin{aligned}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \quad &= \quad -\left[\frac{\partial}{\partial \boldsymbol{\theta}}\,\mathbf{U}(\boldsymbol{\theta})\right]^{-1}\mathbf{U}(\boldsymbol{\theta}) + o_p(1) \\ \sqrt{K}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \quad &\xrightarrow{\text{p}} \quad -\sqrt{K}\left[\frac{\partial}{\partial \boldsymbol{\theta}}\,\mathbf{U}(\boldsymbol{\theta})\right]^{-1}\mathbf{U}(\boldsymbol{\theta})\,.\end{aligned}$$

□

*Corollary (ii).* $E\,[\mathbf{U}_i(\boldsymbol{\theta})] = \mathbf{0}$ and $\text{var}(\,\mathbf{U}_i(\boldsymbol{\theta})\,) = \Gamma_i$. Under certain regularity conditions, by the multivariate central limit theorem, (Serfling (1980), Theorem B, p30),

$$K^{-\frac{1}{2}}\mathbf{U}(\boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \text{MVN}\left(\mathbf{0}, \lim_{K\to\infty} K^{-1}\sum_{i=1}^{K}\Gamma_i\right)\,.$$

□

*Corollary (iii).* Let $E\left[\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta})\right] = \sum_{i=1}^{K} D_i = D$. If the Markov condition (Sen and Singer (1993), Theorem 2.3.7) holds for each element of $\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta})$, then the Markov condition holds for the matrix $\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta})$, and

$$K^{-1}\,\frac{d}{d\boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta}) \xrightarrow{\text{p}} K^{-1}\,D\,.$$

□

## Variance in ORTH association estimating equations

Note that residual $T_{ijk}^{(a.b)}$ is defined for $1 \leq j < k \leq n_i$ and $1 \leq a, b \leq C$ such that

$$T_{ijk}^{(a,b)} = Y_{ij}^{(a)} Y_{ik}^{(b)} - \{ \mu_{ijk}^{(a,b)} + b_{ijk:j}^{(a,b)} (Y_{ij}^{(a)} - \mu_{ij}^{(a)}) + b_{ijk:k}^{(a,b)} (Y_{ik}^{(b)} - \mu_{ik}^{(b)}) \},$$

and $E\left(T_{ijk}^{(a,b)}\right) = 0$. Noting that $E\left(Y_{ij}^{(a)} Y_{ik}^{(b)} Y_{ij}^{(c)} Y_{ik}^{(d)}\right) = E\left(Y_{ij}^{(a \wedge c)} Y_{ik}^{(b \wedge d)}\right) = \mu_{ijk}^{(a \wedge c, b \wedge d)}$, therefore $\mathrm{cov}(\, T_{ijk}^{(a,b)}\, ,\, T_{ijk}^{(c,d)}\,)$ is determined by

$$\begin{aligned}
E\left(T_{ijk}^{(a,b)}\, ,\, T_{ijk}^{(c,d)}\right) &= \left(Y_{ij}^{(a)} Y_{ik}^{(b)} - \{ \mu_{ijk}^{(a,b)} + b_{ijk:j}^{(a,b)} (Y_{ij}^{(a)} - \mu_{ij}^{(a)}) + b_{ijk:k}^{(a,b)} (Y_{ik}^{(b)} - \mu_{ik}^{(b)}) \}\right) \times \\
&\quad \left(Y_{ij}^{(c)} Y_{ik}^{(d)} - \{ \mu_{ijk}^{(c,d)} + b_{ijk:j}^{(c,d)} (Y_{ij}^{(c)} - \mu_{ij}^{(c)}) + b_{ijk:k}^{(c,d)} (Y_{ik}^{(d)} - \mu_{ik}^{(d)}) \}\right) \\
&= E\left(Y_{ij}^{(a)} Y_{ik}^{(b)} Y_{ij}^{(c)} Y_{ik}^{(d)}\right) \\
&\quad - E\left(Y_{ij}^{(a)} Y_{ik}^{(b)} \{ \mu_{ijk}^{(c,d)} + b_{ijk:j}^{(c,d)} (Y_{ij}^{(c)} - \mu_{ij}^{(c)}) + b_{ijk:k}^{(c,d)} (Y_{ik}^{(d)} - \mu_{ik}^{(d)}) \}\right) \\
&\quad - E\left(Y_{ij}^{(c)} Y_{ik}^{(d)} \{ \mu_{ijk}^{(a,b)} + b_{ijk:j}^{(a,b)} (Y_{ij}^{(a)} - \mu_{ij}^{(a)}) + b_{ijk:k}^{(a,b)} (Y_{ik}^{(b)} - \mu_{ik}^{(b)}) \}\right) \\
&\quad + E\left(\{ \mu_{ijk}^{(a,b)} + b_{ijk:j}^{(a,b)} (Y_{ij}^{(a)} - \mu_{ij}^{(a)}) + b_{ijk:k}^{(a,b)} (Y_{ik}^{(b)} - \mu_{ik}^{(b)}) \}\right. \\
&\quad \left. \times \{ \mu_{ijk}^{(c,d)} + b_{ijk:j}^{(c,d)} (Y_{ij}^{(c)} - \mu_{ij}^{(c)}) + b_{ijk:k}^{(c,d)} (Y_{ik}^{(d)} - \mu_{ik}^{(d)}) \}\right) \\
&= \mu_{ijk}^{(a \wedge c, b \wedge d)} \\
&\quad - \mu_{ijk}^{(c,d)} E\left(Y_{ij}^{(a)} Y_{ik}^{(b)}\right) - b_{ijk:j}^{(c,d)} E\left(Y_{ij}^{(a \wedge c)} Y_{ik}^{(b)}\right) + \mu_{ij}^{(c)} b_{ijk:j}^{(c,d)} E\left(Y_{ij}^{(a)} Y_{ik}^{(b)}\right) \\
&\quad - b_{ijk:k}^{(c,d)} E\left(Y_{ij}^{(a)} Y_{ik}^{(b \wedge d)}\right) + \mu_{ik}^{(d)} b_{ijk:k}^{(c,d)} E\left(Y_{ij}^{(a)} Y_{ik}^{(b)}\right) \\
&\quad - \mu_{ijk}^{(a,b)} E\left(Y_{ij}^{(c)} Y_{ik}^{(d)}\right) - b_{ijk:j}^{(a,b)} E\left(Y_{ij}^{(a \wedge c)} Y_{ik}^{(d)}\right) + \mu_{ij}^{(a)} b_{ijk:j}^{(a,b)} E\left(Y_{ij}^{(c)} Y_{ik}^{(d)}\right) \\
&\quad - b_{ijk:k}^{(a,b)} E\left(Y_{ij}^{(c)} Y_{ik}^{(b \wedge d)}\right) + \mu_{ik}^{(b)}) b_{ijk:k}^{(a,b)} E\left(Y_{ij}^{(c)} Y_{ik}^{(d)}\right) \\
&\quad + E\left(\{ \mu_{ijk}^{(a,b)} + b_{ijk:j}^{(a,b)} (Y_{ij}^{(a)} - \mu_{ij}^{(a)}) + b_{ijk:k}^{(a,b)} (Y_{ik}^{(b)} - \mu_{ik}^{(b)}) \}\right. \\
&\quad \left. \times \{ \mu_{ijk}^{(c,d)} + b_{ijk:j}^{(c,d)} (Y_{ij}^{(c)} - \mu_{ij}^{(c)}) + b_{ijk:k}^{(c,d)} (Y_{ik}^{(d)} - \mu_{ik}^{(d)}) \}\right) .
\end{aligned}$$

It can be shown that

$$
\begin{aligned}
E\Big(&\{\,\mu_{ijk}^{(a,b)} + b_{ijk:j}^{(a,b)}(Y_{ij}^{(a)} - \mu_{ij}^{(a)}) + b_{ijk:k}^{(a,b)}(Y_{ik}^{(b)} - \mu_{ik}^{(b)})\,\} \\
&\quad \times \{\,\mu_{ijk}^{(c,d)} + b_{ijk:j}^{(c,d)}(Y_{ij}^{(c)} - \mu_{ij}^{(c)}) + b_{ijk:k}^{(c,d)}(Y_{ik}^{(d)} - \mu_{ik}^{(d)})\,\}\Big) \\
= \;& -(\mu_{ij}^{(a)} b_{ijk:j}^{(a,b)} + \mu_{ik}^{(b)} b_{ijk:k}^{(a,b)})\,(\mu_{ij}^{(c)} b_{ijk:j}^{(c,d)} + \mu_{ik}^{(d)} b_{ijk:k}^{(c,d)}) + \mu_{ijk}^{(a,b)} \mu_{ijk}^{(c,d)} \\
&+ b_{ijk:j}^{(a,b)} b_{ijk:j}^{(c,d)} E\left(Y_{ij}^{(a\wedge c)}\right) + b_{ijk:k}^{(a,b)} b_{ijk:k}^{(c,d)} E\left(Y_{ik}^{(b\wedge d)}\right) \\
&+ b_{ijk:j}^{(a,b)} b_{ijk:k}^{(c,d)} E\left(Y_{ij}^{(a)} Y_{ik}^{(d)}\right) + b_{ijk:j}^{(c,d)} b_{ijk:k}^{(a,b)} E\left(Y_{ij}^{(c)} Y_{ik}^{(b)}\right)\,.
\end{aligned}
$$

Collecting terms and substituting the above, $\mathrm{cov}(\,T_{ijk}^{(a,b)}\,,\,T_{ijk}^{(c,d)}\,)$ simplifies to

$$
\begin{aligned}
&\mu_{ijk}^{(a\wedge c,\, b\wedge d)} - b_{ijk:j}^{(a,b)} \mu_{ijk}^{(a\wedge c,\, d)} - b_{ijk:j}^{(c,d)} \mu_{ijk}^{(a\wedge c,\, b)} - b_{ijk:k}^{(a,b)} \mu_{ijk}^{(c,\, b\wedge d)} \\
&- b_{ijk:k}^{(c,d)} \mu_{ijk}^{(a,\, b\wedge d)} + (\mu_{ij}^{(c)} b_{ijk:j}^{(c,d)} + \mu_{ik}^{(d)} b_{ijk:k}^{(c,d)} - \mu_{ijk}^{(c,d)}) \mu_{ijk}^{(a,b)} \\
&+ (\mu_{ij}^{(a)} b_{ijk:j}^{(a,b)} + \mu_{ik}^{(b)} b_{ijk:k}^{(a,b)} - \mu_{ijk}^{(a,b)}) \mu_{ijk}^{(c,d)} + b_{ijk:j}^{(a,b)} b_{ijk:k}^{(c,d)} \mu_{ijk}^{(a,d)} \\
&+ b_{ijk:j}^{(c,d)} b_{ijk:k}^{(a,b)} \mu_{ijk}^{(c,b)} + b_{ijk:j}^{(a,b)} b_{ijk:j}^{(c,d)} \mu_{ij}^{(a\wedge c)} + b_{ijk:k}^{(a,b)} b_{ijk:k}^{(c,d)} \mu_{ik}^{(b\wedge d)} \\
&- (\mu_{ij}^{(a)} b_{ijk:j}^{(a,b)} + \mu_{ik}^{(b)} b_{ijk:k}^{(a,b)})(\mu_{ij}^{(c)} b_{ijk:j}^{(c,d)} + \mu_{ik}^{(d)} b_{ijk:k}^{(c,d)}) + \mu_{ijk}^{(a,b)} \mu_{ijk}^{(c,d)}\,,
\end{aligned}
$$

the form shown in (4.6).

## ORTH binary equivalence

Non-zero elements of variance matrix $\mathbf{P}_i \approx \mathrm{var}(\mathbf{T}_i)$, for $1 \le a, b, c, d \le C$ have the form given above and in (4.6), for integer valued $C \ge 1$. When outcome $O_{ij}$ is binary, $C = 1$ and the block diagonal matrices in $\mathbf{P}_i$ are scalars. Non-zero elements of $\mathbf{P}_i$ in this case are variances of elements $T_{ijk}$. Also for binary $O_{ij}$ and $O_{ik}$, with $C = 1$,

$a = b = c = d = 1$, reducing (4.6) to

$$\begin{aligned}
\mathrm{var}(\,T_{ijk}\,) &= \mu_{ijk} - b_{ijk:j}\mu_{ijk} - b_{ijk:j}\mu_{ijk} - b_{ijk:k}\mu_{ijk} \\
&\quad -b_{ijk:k}\mu_{ijk} + (\mu_{ij}b_{ijk:j} + \mu_{ik}b_{ijk:k} - \mu_{ijk})\mu_{ijk} \\
&\quad +(\mu_{ij}b_{ijk:j} + \mu_{ik}b_{ijk:k} - \mu_{ijk})\mu_{ijk} + b_{ijk:j}b_{ijk:k}\mu_{ijk} \\
&\quad +b_{ijk:j}b_{ijk:k}\mu_{ijk} + b_{ijk:j}b_{ijk:j}\mu_{ij} + b_{ijk:k}b_{ijk:k}\mu_{ik} \\
&\quad -(\mu_{ij}b_{ijk:j} + \mu_{ik}b_{ijk:k})(\mu_{ij}b_{ijk:j} + \mu_{ik}b_{ijk:k}) + \mu_{ijk}\mu_{ijk} \\
&= \mu_{ijk} - 2b_{ijk:j}\mu_{ijk} - 2b_{ijk:k}\mu_{ijk} \\
&\quad +2(\mu_{ij}b_{ijk:j} + \mu_{ik}b_{ijk:k} - \mu_{ijk})\mu_{ijk} + 2b_{ijk:j}b_{ijk:k}\mu_{ijk} \\
&\quad +b_{ijk:j}^2\mu_{ij} + b_{ijk:k}^2\mu_{ik} - (\mu_{ij}b_{ijk:j} + \mu_{ik}b_{ijk:k})^2 + \mu_{ijk}^2\,,
\end{aligned}$$

which after some manipulation is equivalent to

$$\mathrm{var}(T_{ijk}) = \frac{\mu_{ijk}(\mu_{ij} - \mu_{ijk})(\mu_{ik} - \mu_{ijk})(1 - \mu_{ik} - \mu_{ij} + \mu_{ijk})}{\mu_{ij}\mu_{ik}(1 - \mu_{ij} - \mu_{ik} + 2\mu_{ijk}) - \mu_{ijk}^2}\,.$$

# References


Agresti, A. (1999). Modelling ordered categorical data: recent advances and future challenges. *Statistics in Medicine* **18**, 2191–2207.

Agresti, A. (2003). *Categorical Data Analysis (Second Edition).* Wiley Series in Probability and Statistics.

Agresti, A. and Lang, J. B. (1993). A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika* **80**, 527–534.

Braun, T. M. (2007). A mixed model-based variance estimator for marginal model analyses of cluster randomized trials. *Biometrical Journal* **49**, 394–405.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

By, K., Qaqish, B. F. and Preisser, J. (2008). The orth packageurl: `http://www.unc.edu/~kby/R_Splus/rs.html`.

Carey, V., Zeger, S. L. and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517–526.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34**, 305–334.

Clayton, D. (1992). Repeated ordinal measurements: A generalised estimating equation approachtechnical report, Medical Research Council Biostatistics Unit, Cambridge, U.K.

Crouchley, R. (1995). A random-effects model for ordered categorical data. *Journal of the American Statistical Association* **90**, 489–498.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909–917.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician* **45**, 302–304.

Emrich, L. J. and Piedmonte, M. R. (1992). On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation* **41**, 19.

Evans, B. A., Feng, Z. and Peterson, A. V. (2001). A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance pa-

rameter estimation in longitudinal studies and group randomized trials. *Statistics in Medicine* **20**, 3353–3373.

Ezzet, F. and Whitehead, J. (1991). A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine* **10**, 901–907.

Fay, M. P. and Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* **57**, 1198–1206.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51**, 309–317.

Fitzmaurice, G. M. and Laird, N. M. (2000). Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies . *Biostatistics* **1**, 141–156.

Fitzmaurice, G. M., Lipsitz, S. R., Molenberghs, G. and Ibrahim, J. G. (2001). Bias in estimating association parameters for longitudinal binary responses with drop-outs. *Biometrics* **57**, 15–21.

Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R. (1995). Regression models for longitudinal binary responses with informative drop- outs. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 691–704.

Galecki, A. T., Ten Have, T. R. and Molenberghs, G. (2001). A simple and fast alternative to the EM algorithm for incomplete categorical data and latent class models. *Computational Statistics & Data Analysis* **35**, 265 – 281.

Gange, S. J. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician* **49**, 134–138.

Gange, S. J., Linton, K. L. P., Scott, A. J., Demets, D. L. and Klein, R. (1995). A comparison of methods for correlated ordinal measures with ophthalmic applications. *Statistics in Medicine* **14**, 1961–1974.

Glonek, G. F. V. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 533–546.

Heagerty, P. J. and Zeger, S. L. (1996). Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association* **91**, 1024–1036.

Hogan, J. W., Roy, J. and Korkontzelou, C. (2004). Handling drop-out in longitudinal studies. *Statistics in Medicine* **23**, 1455–1497.

Hughes, G., Cutter, G., Donahue, R., Friedman, G., Hulley, S., Hunkeler, E., Jacobs, D., Liu, K., Orden, S., Pirie, P., Tucker, B. and Wagenknecht, L. (1987). Recruitment in the coronary artery disease risk development in young adults (CARDIA) study. *Controlled Clinical Trials* **8**, 68–73.

Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2006). Analyzing incomplete discrete longitudinal clinical trial data. *Statistical Science* **21**, 52–69.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96**, 1387–1396.

Lesaffre, E. and Molenberghs, G. (1991). Multivariate probit analysis: A neglected procedure in medical statistics. *Statistics in Medicine* **10**, 1391–1403.

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

Liang, K. Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)* **54**, 3–40.

Lin, H., Scharfstein, D. O. and Rosenheck, R. A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **66**, 791–813.

Lipsitz, S. R. and Fitzmaurice, G. M. (1996). Estimating equations for measures of association between repeated binary responses. *Biometrics* **52**, 903–912.

Lipsitz, S. R., Kim, K. and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine* **13**, 1149–1163.

Lipsitz, S. R., Molenberghs, G., Fitzmaurice, G. M. and Ibrahim, J. (2000). GEE with Gaussian estimation of the correlations when data are incomplete. *Biometrics* **56**, 528–536.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* **83**, 1198–1202.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90**, 1112–1121.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data, Second Edition.* John Wiley & Sons, Inc.

Lu, B., Preisser, J. S., Qaqish, B. F., Suchindran, C., Bangdiwala, S. I. and Wolfson, M. (2007). A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics* **63**, 935–941.

Mancl, L. A. and DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57**, 126–134.

Mardia, K. V. (1967). Some contributions to contingency-type bivariate distributions. *Biometrika* **54**, 235–249.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)* **42**, 109–142.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, Inc.

Miller, M. E., Davis, C. S. and Landis, J. R. (1993). The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics* **49**, 1033–1044.

Miller, M. E., Ten Have, T. R., Reboussin, B. A., Lohman, K. K. and Rejeski, W. J. (2001). A marginal model for analyzing discrete outcomes from longitudinal surveys with outcomes subject to multiple-cause nonresponse. *Journal of the American Statistical Association* **96**, 844–857.

MMWR (1994). CDC Surveillance for selected tobacco-use behaviors — United States, 1900–1994. *Morbidity and Mortality Weekly Report* **43**, SS–3.

MMWR (2003). Cigarette Smoking Among Adults — United States, 2001. *Morbidity and Mortality Weekly Report* **52**, 953–956.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate plackett distribution. *Journal of the American Statistical Association* **89**, 633–644.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5**, 99–135.

O'Hara Hines, R. J. , Hines, W. G. S. and Friesen, T. G. (1999). A comparison of two drop-out weighting schemes in the analysis of clustered data with categorical and continuous responses. *Journal of Agricultural, Biological, and Environmental Statistics* **4**, 203–216.

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association* **92**, 1320–1329.

Pan, W. and Wall, M. M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine* **21**, 1429–1441.

Phillips, C., Essick, G., Preisser, J. S., Turvey, T. A., Tucker, M. and Lin, D. (2007). Sensory retraining after orthognathic surgery: Effect on patient perception of altered sensation. *Journal of Oral and Maxillofacial Surgery* **65**, 1162–1173.

Preisser, J. S., By, K., Perin, J. and Qaqish, B. F. (2008). Regression diagnostics for alternating logistic regressions(*submitted)*.

Preisser, J. S., Galecki, A. T., Lohman, K. K. and Wagenknecht, L. E. (2000). Analysis of smoking trends with incomplete longitudinal binary responses. *Journal of the American Statistical Association* **95**, 1021–1031.

Preisser, J. S., Lohman, K. K. and Rathouz, P. J. (2002). Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine* **21**, 3035–3054.

Preisser, J. S., Lu, B. and Qaqish, B. F. (2008). Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations. *Statistics in medicine* **27**, 5764–5785.

Preisser, J. S. and Qaqish, B. F. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika* **83**, 551–562.

Preisser, J. S., Reboussin, B. A., Song, E.-Y. and Wolfson, M. (2007). The importance and role of intracluster correlations in planning cluster trials. *Epidemiology* **18**, 552–560.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825–839.

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* **90**, 455–463.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22**, 300–325.

Qu, A., Lindsay, B. G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823–836.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122–129.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Rotnitzky, A., Robins, J. M. and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics*. Chapman & Hall.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley.

Sharples, K. and Breslow, N. (1992). Regression analysis of correlated binary data: Some small sample results for the estimating equation approach. *Journal of Statistical Computation and Simulation* **42**, 1–20.

Stiger, T. R., Barnhart, H. X. and Williamson, J. M. (1999). Testing proportionality in the proportional odds model fitted with GEE. *Statistics in Medicine* **18**, 1419–1433.

Stokes, M. E., Davis, C. S. and Koch, G. G. (1995). *Categorical Data Analysis using the SAS System*. SAS Institute.

Troxel, A. B. (1998). A comparative analysis of quality of life data from a southwest oncology group randomized trial of advanced colorectal cancer. *Statistics in Medicine* **17**, 767–779.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer Science+Business Media, LLC.

U.S. Bureau of the Census (1998a). U.S. Census Population Survey Rates. Available at `http://www.census.gov/population/socdemo/race/black/tabs97/tab01.txt`.% vspace10pt

U.S. Bureau of the Census (1998b). U.S. Census Population Survey Rates. Available at `http://www.census.gov/prod/3/98pubs/p20-505u.pdf`.

Wagenknecht, L. E., Craven, T. E., Preisser, J. S., Manolio, T. A., Winders, S. and Hulley, S. B. (1998). Ten-year trends in cigarette smoking among young adults, 1986-1996: The CARDIA study. *Annals of Epidemiology,* **8**, 301–307.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models,

and the Gauss–Newton method. *Biometrika* **61**, 439–447.

Williamson, J. and Kim, K. M. (1996). A global odds ratio regression model for bivariate ordered categorical data from ophthalmologic studies. *Statistics in medicine* **15**, 1507–1518.

Williamson, J. M., Kim, K. and Lipsitz, S. R. (1995). Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association* **90**, 1432–1437.

Wolfson, M., Altman, D., DuRant, R., Shrestha, A., Patterson, T. E., Williams, A., Zaccaro, D., Hensberry, R., Suerken, C., Foley, K., Preisser, J. and Brown, S. (2004). National evaluation of the enforcing underage drinking laws program: Year 4 report. Winston-Salem, NC: Wake Forest University School of Medicine, 2004. Available at: `http://www.phsintranet.wfubmc.edu/EUDL2/pubs.cfm`. Accessed February 2, 2007.

Yi, G. Y. and Cook, R. J. (2002). Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association* **97**, 1071–1081.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–648.

Ziegler, A., Kastner, C. and Chang-Claude, J. (2003). Analysis of pregnancy and other factors on detection of human papilloma virus (HPV) infection using weighted estimating equations for follow-up data. *Statistics in Medicine* **22**, 2217–2233.

Zink, R. C. and Qaqish, B. F. (2009). Orthogonalized residuals for estimation of marginally specified association parameters in multivariate binary data. *COBRA Preprint Series 51* URL `http://biostats.bepress.com/cobra/ps/art51`.