View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Scholarship@Western

Western Graduate&PostdoctoralStudies

Western University

Scholarship@Western

Electronic Thesis and Dissertation Repository

4-30-2013 12:00 AM

Statistical Analysis of Correlated Ordinal Data: Application to Cluster Randomization Trials

Ruochu Gao
The University of Western Ontario

Supervisor
Dr. Allan Donner
The University of Western Ontario Joint Supervisor
Dr. Neil Klar
The University of Western Ontario

Graduate Program in Epidemiology and Biostatistics
A thesis submitted in partial fulfillment of the requirements for the degree in Doctor of Philosophy
© Ruochu Gao 2013

Follow this and additional works at: https://ir.lib.uwo.ca/etd

 Part of the Medicine and Health Sciences Commons, and the Physical Sciences and Mathematics Commons

Recommended Citation

Gao, Ruochu, "Statistical Analysis of Correlated Ordinal Data: Application to Cluster Randomization Trials" (2013). *Electronic Thesis and Dissertation Repository*. 1696.
https://ir.lib.uwo.ca/etd/1696

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

STATISTICAL ANLYSIS OF CORRELATED ORDINAL DATA: APPLICATION TO CLUSTER RANDOMZIATION TRIALS

Thesis format: Monograph

by

Ruochu Gao

Graduate Program in Epidemiology & Biostatistics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

© Ruochu Gao 2012

Abstract

Cluster randomization trials have become increasingly popular when theoretical, ethical or practical considerations preclude the use of traditional trials that randomize individual subjects. Although some methods for analyzing clustered ordinal data have been brought to wide attention, these are less developed as compared to methods for analyzing clustered continuous or binary outcome data. The aim of this thesis is to refine existing strategies which may be applicable to clustered ordinal data as well as extensions which have been previously considered only for clustered binary responses. The approaches include adjusted Cochran-Armitage tests using an ICC estimator, and correction and modification strategies to improve the small-sample performance of the Wald test and score test in GEE for clustered ordinal data. The type I error and power for these test statistics are investigated using a simulation study.

Simulation results show that kappa-type estimators had less bias than ICC estimators when cluster sizes were fixed and small for $\rho = 0.005$ or $\rho = 0.01$. Conversely, ANOVA ICCs had relatively smaller bias in the case of variable cluster sizes. In addition, small-sample performance of GEE robust Wald tests are improved by using adjustments and corrections. The adjusted test W_{BC1} is recommended in terms of type I error and power. The discussion is illustrated using data from a school-based cluster randomization trial.

Keywords: cluster randomization; correlated ordinal outcome; ICC estimator; Cochran-Armitage test; GEE; small-sample

Acknowledgments

My deepest appreciation goes to my supervisors, Drs Allan Donner and Neil Klar. I would like to thank them for their persistent patience, warm encouragement and effective guidance while working on the thesis. Their support and insights have been invaluable. This thesis would not have been completed without their continuous help.

I am indebted to Dr. Guangyong Zou and Dr. John Koval for their excellent lectures in Biostatistics and enthusiastic help.

With all my affection, I thank my parents, my husband and my son, for their unfailing love and support. To my family I dedicate this thesis.

Table of Contents

List of Tables

Chapter 1

1 Introduction

1.1 Cluster Randomization Trials

When allocation of individual participants is possible, the randomized clinical trial is generally regarded as the gold standard for the evaluation of interventions in health research. Over the past two decades, random assignment at higher levels of aggregation has become increasingly popular when theoretical, ethical or practical considerations preclude the use of traditional trials that randomize individual subjects (Donner and Klar, 2000, pp. 5). Trials which assign interventions at higher levels of aggregation are referred to as cluster randomization trials. The units of randomization may be families, classrooms, worksites, hospitals or communities.

The reasons for adopting cluster randomization are various, including greater administrative efficiency and the possibility of less experimental contamination (Donner and Klar, 2000; p2-4). There is also, at times, no alternative to cluster randomization as for community intervention trials when the intervention is delivered at the community level, e.g. intervention programmes that use mass media to promote smoking cessation. Gail et al. (1992), for instance, designed the COMMIT (Community Intervention Trial for Smoking Cessation) to study public education and media campaign programmes to accelerate smoking cessation among heavy smokers and to reduce smoking prevalence. As discussed by Gail et al. (1992), these community-based interventions have the potential to affect every smoker in the community. Thus, the intervention precluded individual randomization within communities.

An important feature of cluster randomization trials is that responses of subjects from the same cluster tend to be more alike than responses of subjects from different clusters and thus they are not statistically independent (i.e. are correlated). Within-cluster similarities in response lead to a reduction in effective sample size, and consequently ignoring clustering at the design stage may lead to an underpowered study and a loss of precision for estimating the intervention effect (Donner and Klar, 2000; p6). Furthermore, the

confidence interval for the estimated intervention effect will be too narrow and could lead to a spuriously statistically significant test result. Therefore, the correlation among responses of individuals in the same cluster must be taken into account in both the design and the statistical analysis.

A review conducted more than a decade ago (Simpson et al., 1995) found that design and analysis issues associated with cluster randomization trials were not recognized widely enough. They found that only 4 of 21 trials they reviewed accounted for between-cluster variability in sample size or power calculations, and 12 of 21 trials took account of the effect of clustering in the analysis. Although the number of published randomization trials continues to increase, Varnell et al. (2004) reported that there has been little improvement in the quality of reporting cluster randomization trials from 1998 through 2002.

Fortunately, a recent review (Eldridge et al., 2008) suggests that there has been considerable improvement in the reported design and analysis of cluster randomization trials in primary care trials. Eldridge et al. (2008) reported that 21 of 34 trials they reviewed accounted for clustering in sample size calculations, and 30 of 34 trials took account of clustering effects in analysis. However, this progress is not universal. For instance, Murray et al. (2008) reviewed 75 articles describing applications of cluster randomization trials to cancer research in 41 journals from 2002 to 2006. They reported that only 45 percent of the articles used the appropriative methods to analyze the results.

1.2 Scales of measurements

Steven (1946) defined measurement as “the assignment of numerals to objects or events according to rules”. He proposed four scales of measurement: ratio, interval, ordinal and nominal.

Outcomes measured on ratio and interval measurement scales are typically continuous. Differences between numeric values are meaningful for both ratio and interval measurements. Ratio scale measurements have the additional property of a meaningful zero score indicating the absence of the quantity being measured (Porta et al., 2008).

Donner and Klar (2000) provide examples of analyses where the study outcomes in cluster randomization trials are continuous and measured on a ratio scale. For example, change in cholesterol level (mmol/L) was the primary endpoint measured on students who participated in the Child and Adolescent Trial for Cardiovascular Health (CATCH) – a school randomized trial (Luepker et al., 1996).

Outcomes measured on an ordinal scale may be classified into ordered qualitative categories. However the interval between ordered categories is typically unknown and possibly unmeasurable for ordinal scale outcomes thus distinguishing them from interval and ratio scale measurements. An example is provided by Kim et al., (2005) in their cluster randomization trial which evaluated treatment of rheumatoid arthritis using an adjectival scale (Streiner and Norman, 2003 pp. 33-35). The outcome of interest was patient self-assessment of their attitude classified into three categories: poor, fair or good.

Data measured on a nominal scale are unordered and thus only gives identification values or labels to various categories. Objects with the same value are the same on some attribute or attributes. The values of the scale have no 'numeric' meaning in the way that one usually thinks about numbers. Cook and Demets (2008) observed that randomized trials rarely have nominal categorical outcomes with three or more levels. They noted that "an unordered categorical variable with three or more levels is usually not a suitable outcome measure because there is no clear way to decide if one treatment is superior to another". Binary data are a special case of nominal data with only two categories. An example of binary data is provided by Murray et al. (1992) in their study to evaluate the effect of school-based interventions in reducing adolescent tobacco use. One of the outcomes was if students reported using smokeless tobacco or not.

1.3 Ordinal outcome data

In this thesis attention is limited to analyses of ordinal data obtained from cluster randomization trials.

1.3.1 Number of categories

Ordinal endpoints for randomized trials often use health measurement scales. One should then limit attention to scales which have had their psychometric properties validated. Even then there may be more than one possible choice of scale. The decision as to which scale should be selected as the endpoint will depend, in part, on the number of ordinal categories.

Suppose it is reasonable that the ordinal outcome measures some underlying continuous psychological construct (e.g. pain). Then selection of a more finely graded outcome should increase power to detect an intervention effect to the extent that subjects can discriminate between categories. In practice, there is likely little gain in power by increasing the number of categories beyond about five. This may reflect, in part, the difficulty people have in classifying objects or experiences into much more than seven levels (Schaeffer and Presser, 2003; Streiner and Norman, 2003, p28-29).

Decisions about the number of categories also have implications for data analysis. For example, the weighted kappa statistic varies as a function of category number (Brenner and Kliebsch, 1996).

1.3.2 The Television, School, and Family Smoking Prevention and Cessation Project

The primary outcome in most cluster randomization trials is binary or quantitative (Donner and Klar, 2000; p128). However ordinal data have also been used in a number of cluster randomization trials. Examples of such trials are provided in Table 1.1.

Table 1.1: Examples of recent cluster randomization trials with ordinal outcomes

Reference	Cluster	Outcome	Levels of outcome	Number of Levels
Flay et al. 1995	school	smoking intention	increased, no change, or decreased	3
Marinacci et al. 2001	school	frequency of condom use	always, often or sometimes, never	3
Patton et al. 2006	school	antisocial behavior in the past 6 months	none, once, more than once	3
		tobacco use in past month	none, once to three times, more than three times	3
Glasgow et al 2005	general practitioner	patient satisfaction	yes, doubtful or no	3
Byng et al. 2004	medical practices	severity of mental illness	none, mild, moderate, or severe	4
Klepp et al. 1997	school	communication with AIDS in the past month	never to more than 4 times	4
McCusker et al.1992	medical practices	drug-use behavior	no injection, injection but no borrowing, borrowing but bleach always used, bleach used sometimes, bleach never used	5
Howard-Pitnet et al. 1997	class	nutritional attitude	strongly agree to strongly disagree	5
Rosendal et al. 2003	physicians	classification of the patient problem	physical disease, probable physical disease, medically unexplained symptoms mental illness, no physical symptoms	5
Seligman et al. 2005	physicians	physician satisfaction	very dissatisfied to very satisfied	6
Watson et al. 2005	families	severity of injury	minor, moderate, serious, severe, critical, or unsurvivable	6

Flay et al. (1995) report on a school-based smoking prevention programme. Seventh-grade students were randomized by school into a school-based social resistance curriculum or a television-based tobacco use prevention and cessation programme using a factorial design. Study outcomes of interest included measures of tobacco and health knowledge, coping skills and the prevalence of tobacco use. There were 7351 students who participated in the pretest assessment. These students came from 340 classrooms drawn from 47 schools.

Study outcomes included a tobacco and health knowledge scale defined as the number of correct answers to seven questions. Hedeker and Gibbons (1996) described application of a mixed effects ordinal logistic regression model to examine the effect of intervention on tobacco and health knowledge. For these analyses outcomes were grouped into quartiles given by 0-1, 2, 3 and 4-7 correct answers. These data will be used to illustrate methods of analysis for correlated ordinal outcomes. Detailed analyses are provided in Chapter 7.

1.4 Analysis of Independent Ordinal Outcomes

1.4.1 Overview of Statistical Approaches

Analytic methods for clustered ordinal outcomes are largely extensions of analytic methods for independent ordinal outcomes. Methods for analysis of independent ordinal outcome data may be classified into three approaches: non-parametric, simple linear regression and ordinal logistic regression. Moreover, attention is restricted to methods comparing two independent samples. Additionally, a distinct classification is provided by Agresti and Coull (2002) where they distinguish methods for clustered ordinal outcomes by inequality constraints. However, they noted that inequality-constrained methods are not prominent in the literature and software used for data analysis.

Non-parametric methods may be preferred for testing the effect of intervention when the assumption of normality is questionable. Corresponding two sample approaches include the sign test, the Mann-Whitney-U test, and the Wilcoxon rank sum test. Note that the Wilcoxon rank sum test is equivalent to the Mann-Whitney-U test.

Another common strategy for ordinal data analysis is the assignment of scores to categories and then simply treating the scores as continuous and fitting these using linear models which assume outcomes are normally distributed. This approach has the virtue of familiarity, yielding easily interpretable, albeit potentially misleading, coefficients; however, limitations may arise from ignoring either the discrete nature or the potentially skewed distribution of ordinal data, thus violating the normality assumption. When models for continuous data are directly applied to ordinal data, a further problem is that the ceiling and floor effects of the dependent variable can result in biased estimates of the regression coefficients (McKelvey, 1975; Hedeker and Gibbons, 1994). The robustness and power of this strategy were investigated through computer simulation by Sullivan and D'Agostino (2003). Interestingly, the type I error rates obtained for tests of the effect of intervention were at the nominal level when two sample t-tests were used and when an analysis of covariance (ANCOVA) model with a common slope was fit.

The ordinal nature of the study data may be more appropriately accounted for using generalized linear models (GLM). A popular model for ordinal outcome data is the proportional odds model using cumulative logits (McCullagh, 1980; Hosmer and Lemeshow, 2000, pp. 297), which assumes identical proportionality for each logit (Agresti, 2001). This model is also called the cumulative logit model. In contrast, non-proportional odds ratio extensions of this model permit a separate effect for each logit (Peterson & Harrell, 1990; Agresti, 2001). In addition to the logit link, other link functions possible for ordinal data include the probit link and the complementary log-log link (McCullagh, 1980). These are not discussed further as they are not commonly applied to analyses of epidemiologic data.

When the cumulative logit models fit poorly, one may alternatively fit adjacent-category logits or continuation-ratio logits for ordinal data. Note that the adjacent-category logit model is a special case of the baseline-category logit model which is commonly used for nominal outcome data as a polytomous logistic regression. Liu and Agresti (2005) reviwed recent developments of analysis for ordinal outcomes and reported that the most popular model for ordinal responses uses logits of cumulative probabilities.

Statistical inferences may be conducted using Wald, score or likelihood-ratio methods. The Wald test uses information from the curvature of the log-likelihood function and the distance between the parameter estimate and the null parameter value. The score test is based on the slope and curvature of the log-likelihood function only at the null parameter value. It does not require the computation of a parameter estimate. The likelihood-ratio test combines the information about the log-likelihood function at both the null value and estimated value of the parameter. Hauck and Donner (1977) showed the Wald tests for coefficients from a logistic regression model may behave in an aberrant manner in that power can decrease even as the estimated regression coefficient gets larger. This behavior of Wald tests is particularly likely when the sample size is small. They recommended that the likelihood ratio test be used instead.

These model-based tests are often equivalent, at least in special cases, to well known non-parametric test statistics. For instance, the score test from a proportional odds model for a two-group comparison is identical to the Wilcoxon rank sum test (McCullagh, 1980). However, some adjustment for these tests will be needed when applied to clustered ordinal data.

1.4.2 Scoring Ordinal Outcomes

The statistical methods which have been reviewed may be distinguished by the method used to account for the inherent order of the categories or equivalently by the choice of inequality constraint (Agrest and Coull, 2002). This is accomplished for non-parametric and parametric methods by imposing a scoring scheme to the qualitative ordered categories while ordinal logistic regression accounts for ordinality by imposing constraints on the odds ratios.

Some authors have argued for application of non-parametric methods based on the false assumption that it is then not necessary to impose an arbitrary choice of score (Graubard and Korn, 1987). This overstates the situation as non-parametric methods score qualitative ordered categories using mid-ranks – a function of the data.

Methods of assigning scores have been described by Armitage (1955) and by Graubard and Korn (1987):

1. Scores may be linearly related to a quantitative measurement when the ordinal outcome is obtained by degrading a variable which is more finely measured, e.g. using midpoints of categories formed by grouping scores from a health measurement scale into quartiles.

2. When no natural category scores are available equally spaced scores are often selected to detect linear components of the intervention effect although rank scores may then also be used.

3. Sensitivity analysis is recommended to explore the effect of scores on study conclusions.

Furthermore, Kimeldorf et al. (1992) reviewed the statistical tests to compare ordinal outcomes from two samples and the scores adapted for each test. They further proposed an approach to obtain the minimum and maximum values of these test statistics over all possible assignments of scores. Thus if the range of the minimum and maximum values includes the critical value of the tests statistic, they suggested that one must be aware to justify the choice of scores used in the analysis.

In this thesis, I will consider the effect of score choice on validity and power of extensions of the Cochran-Armitage test which adjusts for clustering. Additionally I will compare the Cochran-Armitage test statistic to the Wilcoxon rank sum test exploring relationships between these methods.

1.5 Analysis of Clustered Ordinal Outcomes

1.5.1 Overview of Statistical Approaches

The degree of similarity among responses within a cluster is typically measured by the intracluster correlation coefficient (ICC). Denoted by the Greek letter ρ, it may be interpreted as the proportion of overall variation in responses that can be accounted for by between-cluster variation. A more comprehensive measure of the effect of clustering is

given by the design effect $DE = 1+(m-1)\rho$. This parameter measures the amount by which one must increase a standard variance estimate to allow for clustering, and therefore also is often referred to as the variance inflation factor. One can use design effects to adjust standard statistical approaches for clustered data at both the design and analysis stage (e.g., Donner and Donald, 1988). One advantage of this relatively simple approach is that it avoids intensive computation. Additionally, Scott and Holt (1982) derived a design effect for the variance of estimated regression coefficients from a linear regression model, while Neuhaus and Segal (1993) extended this result to logistic regression.

The unit of analysis for clustered data may be at either the cluster level or the individual level. Schools were randomly assigned to the intervention groups as part of the Child and Adolescent Trial for Cardiovascular Health (CATCH, Zucker at al, 1995). One of the secondary objectives of this trial was to evaluate the effect of training food service personnel on the dietary quality of food services (e.g., to decrease fat content). This outcome variable was collected from school lunch menus and thus the analysis was necessarily conducted at the school level. On the other hand, health outcomes analyses were conducted at the individual level. Advantages of cluster-level analyses include the possible construction of exact statistical inferences and valid tests of significance when there are small numbers of clusters; whereas individual-level analysis allows direct examination of cluster-level and individual-level predictors and provides more efficient estimates of the effect of intervention when cluster sizes are variable, assuming adjustment for effects of clustering (Donner and Klar, 2000; p80). A unique challenge for ordinal outcomes, however, is the specification of an appropriate cluster-level summary statistic. Because of this challenge we limit attention to individual level analyses.

Some standard non-parametric methods have been extended to the case of clustered ordinal outcome data. Rosner et al. (2003 and 2006), Rosner and Grove (1999) and Brunner and Langer (2000) extended the Wilcoxon rank sum test and the Wilcoxon signed rank test to clustered data. Furthermore Jung and Kang (2001) derived a test statistic unifying the Wilcoxon rank sum test and the Cochran-Armitage trend test for clustered ordinal data.

Modeling approaches have also been extended to correlated ordinal data. As in the case of independent ordinal data, an approach for analyzing clustered ordinal data is to treat ordinal responses as continuous and then apply more familiar approaches to the clustered continuous responses. However, Hedeker and Gibbons (1994) claimed that this strategy could bias estimated regression coefficients due to the floor and ceiling effects of outcomes. Moreover, Fielding et al. (2003) compared parameter estimates obtained using multilevel linear models and multilevel ordinal models by analyzing data on educational examination grades. They reported that the magnitude and precision of fixed effect estimates were quite similar between the two models. However, random effect estimates with continuous outcomes are somewhat sensitive to the choice of score and their precision differs from that of ordinal models. These differences need to be further examined using simulation.

Extensions of generalized linear models for analysis of correlated data may be classified as population-average models (e.g., marginal model), cluster-specific models (e.g., generalized linear mixed models) or transition models. Discussions of these models for binary data include Diggle et al. (1994), Pendergast et al. (1996), and Heagerty and Zeger (2000). In addition, Agresti and Natarajan (2001) provided a comprehensive review of marginal and cluster-specific models for ordinal outcome data.

Generally, transition models focus on the dependence of a response on previously observed responses and treat them as explanatory variables of the current response. So they are always used for repeated measurement analysis. Thus transition modeling methods will not be considered in this study.

On the other hand, cluster-specific models focus on cluster-level effects while marginal models emphasize the average effect at the population level. Therefore, marginal models are more relevant in analyses of data arising from cluster randomization trials than cluster-specific models. Particularly, our interest is on the intervention effect on average population level. Hence, more attention is given to marginal models than cluster-specific models in this study.

In addition, there has been considerable attention given to their limitations. Agresti and Natarajan (2001), for instance, noted maximum likelihood fitting methods require intensive computation. Other marginal modeling strategies, for instance, Dirichlet-multinomial modeling method, could reduce the required intensive computation because the number of parameters does not vary with cluster size. As an alterative, the generalized estimating equation (GEE) approach requires specification of only the first two moments but the associated robust variance estimator is biased downward when there are few clusters (e.g., Murray et al., 2004). For cluster-specific models, maximum likelihood become challenging when there are more than five random effects. In particular, the use of Gauss-Hermite quadrature approach for approximating the likelihood function will be limited (Hedeker, 2003). However, admittedly the challenge to fitting mixed effects models noted by Hedeker (2003) is of limited concern for most cluster randomization trials as then concern typically focuses on only a single between-cluster source of random variation.

1.5.2 Estimation of the Intracluster Correlation Coefficient (ICC)

1.5.2.1 Estimation for Clustered Continuous and Binary Data

Various estimators of the ICC have been reviewed in the literature (Donner, 1986; Ridout et al., 1999). There are at least three frequently used estimators of the ICC for clustered continuous and binary data. These include the one-way analysis of variance (ANOVA) estimator, the method of moments estimator and the fully parametric approach estimator. Klar (1993, p57-61) gave detailed discussions on these three methods for clustered binary outcomes.

1.5.2.2 Estimation for Clustered Ordinal Data

Approaches for estimating the ICC for continuous and binary data could be extended to clustered ordinal data. For example, the ANOVA methods could be used for clustered ordinal data by assigning scores to ordered categories. Moment-based methods, such as estimator obtained from marginal proportional odds logistic models using the GEE approach (Liptisz et al., 1994) have also been proposed. Additionally, one could estimate

the ICC for clustered ordinal data by assuming that the study outcome follows a Dirichlet-multinomial distribution (Lui et al., 1999).

1.5.3 Non-parametric Approaches

Non-parametric methods occupy an important role given that they perform well without the need to make distributional assumptions. Simple adjustments to standard methods also allow them to be applied to clustered ordinal data. For example, Rosner and Grove (1999) generalized the Wilcoxon rank sum test to account for clustering by introducing four separate correlation parameters into the variance formula; Brunner and Langer (2000) extended the same test by formulating nonparametric hypotheses by means of the marginal distribution of treatment effects. Rosner et al. (2003 and 2006) generalized variance formulae for the Wilcoxon rank sum test and the Wilcoxon signed rank test that account for clustering effects.

1.5.4 Marginal Models

1.5.4.1 Maximum likelihood (ML) fitting

The likelihood function for a marginal logit model may be constructed as the multinomial joint probabilities while the marginal model refers to marginal probabilities. Thus it may involve complicated computation to fit marginal models using ML directly. One approach treats the model as a set of constraints on the cell probabilities and then maximizes the likelihood subject to these constraints (Lang and Agresti, 1994). This method is also referred as Lagrange's method (Aitchison and Silvey, 1988; Haber, 1985; Haber and Brown, 1986). As such the marginal model could be equivalently expressed as the constraint model, and Haber (1985) used a Newton-Raphson algorithm to maximize the corresponding Lagrangian likelihood equation. Additionally, Glonek and McCullagh (1995) and Glonek (1996) presented a one-to-one correspondence between joint probabilities and a loglinear model that is composed of marginal probabilities and higher-order loglinear parameters. The likelihood is then maximized in terms of the two sets of models. One is the model specified for the marginal probabilities and the other is the one specified for the high-order parameters. Agresti and Natarajan (2001) provide a detailed review of maximum likelihood approaches under marginal models.

In addition, the Dirichlet-multinomial distribution has been used to model clustered ordinal outcome data. For example, Chen and Li (1994) proposed a quasi-likelihood approach to model the association between two proportions under Dirichlet-multinomial distributions, and Lui et al. (1999) described interval estimators for the ICC and odds ratio for this model.

1.5.4.2 Generalized Estimation Equation (GEE) Approach

Lipsitz et al. (1994a) extended GEE methodology (Liang and Zeger, 1986) to marginal modeling with an ordinal response. Using a multivariate generalization of quasi-likelihood, the GEE regression estimators are consistent and the covariance estimators which use the sandwich form (e.g. sandwich estimator) are robust even with misspecification of the assumed covariance structure (Liang and Zeger, 1986). Statistical inferences may be accomplished using Wald or score test statistics. An alterative to the sandwich estimator is a model-based variance estimator, which is based on the assumed covariance structure. The sandwich estimator uses empirical evidence from the data to adjust the model-based variance in case the assumed covariance structure differs from the true one.

In spite of the wide use of GEE, small-sample performances of sandwich (robust) variance estimators for binary data have been investigated (e.g., Kauermann and Carroll, 2001; Feng and Braun, 2002). In particular, simulation studies show that the sandwich variance estimator tends to underestimate the true variance when the number of clusters is less than 50 (e.g., Mancl and DeRouen, 2001). Consequently, the type I error for the Wald chi-square test using the sandwich estimator is inflated and the resulting confidence interval tends to be too narrow. In contrast to the liberal behaviors of robust Wald tests, Guo et al. (2005) reported that in this case the robust score test using the sandwich estimators has smaller test sizes than the nominal level.

1.5.4.3 Sandwich Estimator Corrections for Clustered Binary Data

A variety of small-sample adjustments and modifications for the sandwich variance estimator have been proposed and compared. Mancl and DeRouen (2001) applied the Student's t- or F-distribution instead of the normal or chi-square distribution for

significance testing. Lipsitz et al. (1994b) recommended using the one step GEE estimators instead of the fully iterated estimators when the binary responses are highly correlated. Additionally, resampling methods, such as the jackknife and bootstrap, have also been considered (Lipsitz et al. 1994; Sherman and le Cessie, 1997; Feng et al. 1996).

In the sandwich estimator, the unknown covariance matrix is estimated by residuals. When the number of clusters is small, the residuals tend to be negatively biased leading to underestimation of the covariance matrix. Mancl and DeRouen (2001) proposed a bias-corrected sandwich estimator by modifying the residual.

In addition, when the number of clusters is small, standard normal critical values are no larger appropriate. Kauermann and Carroll (2001) used a function of the variance of the sandwich estimator to adjust the normal distribution quantiles. Pan and Wall (2002) proposed a more general approach, adjusting the approximate t- or F-test by the variability of the sandwich estimator.

1.5.5 Cluster-specific Models

Cluster-specific models represent an extension of the generalized linear model that permits random effects as well as fixed effects (Agresti, 2002). The inclusion of random effects allows specification of the correlation between observations within a cluster.

The likelihood is a function of the marginal distribution obtained after integrating out the unobservable random effects. This integral rarely has a closed form and therefore it is necessary to approximate the likelihood function. Hedeker and Gibbons (1994) derived the Gauss-Hermite quadrature approximating the integral by a weighted sum at certain points. In order to increase its efficiency, Liu and Pierce (1994) proposed an adaptive version of Gauss-Hermite quadrature. As an alternative, the quasi-likelihood method, one of the Laplace approximation methods, avoids the integration problem and is feasible for large data sets (Breslow and Clayton, 1993). However, it performs poorly when the variance components are large (McCulloch, 1997). Other approaches for approximating the integration over the random effects include Gibbs sampling (Zeger and Karim, 1991), a combination of Monte Carlo with Newton-Raphson (McCulloch, 1997) or EM

algorithm (Booth and Hobert, 1999), and simulating the likelihood function directly by MCMC (McCulloch, 1997).

Instead of assuming a parametric distribution for the random effects, Aitkin (1999) and Santos and Berridge (2000) proposed a non-parametric mixing distribution to specify the distribution of random effects, as approximated by some mass points. Hartzel et al. (2001) combined this with the EM algorithm. Morel and Nagaraj (1993) further proposed a finite mixture distribution to model clustered categorical data.

1.6 Testing Assumptions of Ordinal Outcome Data

Armitage (1955) and Cochran (1954) derived two chi-square test statistics: one assesses the deviations from linearity of the outcome data, and the other tests the trend among binomial proportions of ordered groups. The statistic for testing the deviation from linearity could be obtained from the difference between the Pearson test statistic for association .and the trend test statistic. An analogous examination of ordinality was considered by Imrey et al. (1981) and Brant (1990) in the context of assessing assumptions of proportionality for ordinal logistic regression models.

For clustered data, Donner and Donald (1988) derived an adjusted Pearson chi-square test and an adjusted chi-square trend test. Consequently, both the Pearson and trend test statistics have been extended for clustered data. However, whether the statistic testing ordinality of clustered ordinal data could be simply obtained from the difference between those two statistics is a future topic. An analogous examination of ordinality is extending assumption assessment in ordinal regression model for independent data to clustered data. For example, Stiger et al. (1999) considered both a score test and a Wald test for assessing the assumption of proportional odds in the proportional odds model fitted with GEE.

1.7 Scope of the Thesis

Cluster randomization trials are often distinguished by the size of the unit randomized. Trials randomizing small units (e.g. families) typically enroll large numbers of such clusters. Conversely economic and practical constraints typically limit the number of

clusters recruited to community intervention trials. In this thesis I limit attention to community intervention trials since these tend to have greater statistical challenges. For example, the validity of statistical inferences is often problematic when there are few large clusters.

The number of ordinal categories used in most practical applications ranges from three to five (Brenner and Kliebsch, 1996). In this thesis we restrict our attention to ordinal data with three categories. Extension of all methods to more categories is straightforward.

There are three designs that are most frequently used in cluster randomization trials: completely randomized, matched-pair and stratified. The completely randomized design is suited to trials that have a fairly large number of clusters; whereas matching or stratification is more desirable in studies with few clusters. Furthermore regression models described in this proposal may be directly extended to the stratified design. The challenge of extending the methods discussed here to pair-matched designs poses problems that are an area for future research and will not be discussed further here. The discussion will also be focused on models where there is a single binary, cluster-level covariate, i.e., trials where there is one experimental and one control group.

Among the methods reviewed above, the primary emphasis of my research is on non-parametric methods, marginal modeling, and cluster-specific modeling as applied to clustered ordinal data. For model-based methods, we limit attention to cumulative logit links.

1.8 Objectives

Only limited research has been carried out exploring the unique challenges of analyzing ordinal outcome data arising from cluster randomization trials. The principle challenge is that methods need to account for dependencies in outcome among cluster members. Although methods for analyzing clustered ordinal data were brought to wide attention in the last two decades, such methods are not as developed as methods for analyzing clustered continuous or binary outcome data. In this research, I will highlight refinements

of existing strategies which may be applicable to clustered ordinal data as well as extensions which have been previously considered only for clustered binary responses.

Analytically, I will formulate a Cochran-Armitage test statistic for clustered ordinal outcomes data estimating an intracluster correlation coefficient for correlated ordinal data. This approach does not require complex computation or software proposed by other methods. In addition, I will develop some correction and modification strategies to improve the small-sample performance of the Wald test and score test in GEE for clustered ordinal data.

In addition to this analytic work, I will conduct simulation studies comparing the performance of model-based methods on bias and standard errors of estimators as well as type I error and statistical power. Furthermore, I will evaluate the small-sample performance of the score and Wald tests applied in GEE for clustered ordinal outcome data. To improve their performance, I will extend small-sample adjustments proposed for the sandwich variance estimators to clustered ordinal outcome data and present a comparison of their properties.

Finally I will use data from the Television, School, and Family Smoking Prevention and Cessation Project (TVSFP) to illustrate results. From the literature review, data from the TVSFP have been widely used as examples in studies involving clustered ordinal outcome data. Hedeker et al. (1994), for instance, analyzed data from the TVSFP by using a linear random effects model; and Hedeker and Gibbons (1994), Sashegyi et al. (2000), and Raman and Hedeker (2005) analyzed it by using ordinal random effects models. In addition, Yang (2001, pp. 107-125) and Fitzmaurice et al. (2004, pp. 5) used it to illustrate methods in their books.

Chapter 2

2 Estimating Intracluster Correlation Coefficient

2.1 Introduction

One of the defining features of a cluster randomization trial is the similarity among responses within a cluster, which is measured by the intracluster correlation coefficient ρ. To discuss methods of analysis for clustered data, the natural starting point is the estimation of the intracluster correlation coefficient (ICC).

Various estimations of the ICC for clustered continuous and binary outcome data have been proposed, as reviewed by Donner (1986) and Ridout et al. (1999). One could extend methods for estimating the ICC for clustered continuous and binary outcomes to ordinal outcomes. For example, Lipsitz et al. (1994) extended Liang and Zeger's (1986) GEE approach to the proportional odds model for ordinal outcome data and proposed a moment ICC estimator. Moreover, Lui et al. (1999) generalized numerous early works (Tamura and Young, 1987; Elston, 1977; Yamamoto and Tanagimoto, 1992) and derived stabilized moment estimator, the "unbiased" moment ICC estimator, and the ANOVA estimator under a Dirichlet-multinomial model.

A simulation study conducted by Ridout et al. (1999) examined 20 different ICC estimators for clustered binary outcomes and identified the ANOVA ICC estimator as one of the three most accurate estimators with respect to both the bias and the mean square error. Moreover, Yamamoto and Yanagimoto (1992) compared the ANOVA ICC estimator for binary data with the MLE estimator, the moment estimator, the 'unbiased' estimator, and the stabilized estimator under a beta-binomial model. They reported that the ANOVA estimator is generally preferable to the MLE and other moment estimators in terms of the bias and mean squared error. Additionally, Donner and Donald (1988) compare the ANOVA estimator with the moment estimator for their uses in their adjusted Pearson chi-square test for clustered binary data. Simulation results show that the former tends to be consistently more accurate than the latter with respect to mean squared error.

In addition to binary outcome data, the ANOVA estimator is frequently used for clustered ordinal outcome data by assigning scores to ordered categories. For instance, Lui et al. (1999) and Lui (2002) derived interval estimators of the ICC and the odds ratio for clustered ordinal outcomes by using the ANOVA ICC estimator under Dirichlet-multinomial distribution. The virtues of using the ANOVA estimator also include that it does not require any specialized software and sophisticated numerical procedure as other model-based approaches do (e.g., the GEE procedure). Thus, these findings and favourable properties lead us to consider using the ANOVA estimator to measure the ICC for clustered ordinal outcome data in our research.

In addition, the estimation of the ICC for clustered outcomes could arise from the literature on the close relationship between measures of intracluster correlation and interobserver agreement. Fleiss and Cuzick (1979) developed a kappa-type ICC estimator for correlated binary outcome data using direct probability calculation. Ridout et al. (1999) reported that the kappa-type ICC estimator by Fleiss and Cuzick (1979) and the ANOVA estimator are two of the three most accurate ICC estimators in terms of bias and mean square errors. Moreover, Mak (1988) proposed another kappa-type ICC estimator and noted that his kappa-type ICC estimator may yield higher efficiency than the ANOVA estimator when ρ is not close to zero. In this chapter we will propose a kappa-type ICC estimator for clustered ordinal outcome data.

The remainder of the chapter is organized as follows. Section 2.2 gives notations used in this thesis. Section 2.3 gives a detailed description of the ANOVA ICC estimator and then briefly introduces other ICC estimators for clustered ordinal data. In section 2.4 we propose a kappa-type ICC estimator and explore its properties and relationships with the ANOVA estimator. In section 2.5 we summary the ICC estimators presented here in a table.

2.2 Notations

To establish notations, consider a cluster randomization trial in which n_i clusters are randomly assigned to each of the treatment group and control group ($i = 1$ or 2). We suppose there are m_{ij} observations in the ijth cluster ($j = 1,2, \ldots, n_i$). Outcomes for each

observation may be classified into one of K ordinal categories. Let Y_{ijlk} = 1 if the lth observation in the jth cluster from the ith group falling into the kth category and 0 otherwise, $l = 1,2, \ldots, m_{ij}$, $k = 1,2, \ldots, K$.

We also use the following notations throughout this thesis:

$M = \sum_{i=1}^{2} \sum_{j=1}^{n_i} m_{ij}$, the total number of observations

$N = \sum_{i=1}^{2} n_i$, the total number of clusters

$M_i = \sum_{j=1}^{n_i} m_{ij}$, the total number of observations in the ith group

$Y_{ijl} = S_k$, the assigned score of the kth category in which the ijlth observation falls

Y_{ijlk}, the indicator of outcomes where Y_{ijlk} = 1 if Y_{ijl} fall in the kth category and 0 otherwise

$Y_{ijk} = \sum_{l=1}^{m_{ij}} Y_{ijlk}$, the number of observations in the ijth cluster falling into the kth category

$Y_{ik} = \sum_{j=1}^{n_i} \sum_{l=1}^{m_{ij}} Y_{ijlk}$, the number of observations from the ith group in the category k

$Y_k = \sum_{i=1}^{2} \sum_{j=1}^{n_i} \sum_{l=1}^{m_{ij}} Y_{ijlk}$, the number of observations falling in category k

$\overline{Y_{ij}} = \sum_{l=1}^{m_{ij}} Y_{ijl} / m_{ij}$, the numerical mean in the ijth cluster

$\overline{Y_i} = \sum_{j=1}^{n_i} \sum_{l=1}^{m_{ij}} Y_{ijl} / n_i$, the numerical mean in the ith group

$\overline{Y} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} \sum_{l=1}^{m_{ij}} Y_{ijl} / N$, the mean scores all over clusters

S_k, the score associated with the category k.

2.3 Methods of Estimation

Techniques of the ICC estimation for ordinal outcome data have been less well developed since estimations of the ICC for clustered ordinal data are not as straightforward as those for clustered continuous and binary outcome data. One of the challenges is to define a method of describing the ordinality.

One of the commonly used methods for dealing with ordinality is to assign scores to ordinal categories. For instance, the moment-based estimators Lui et al. (1999) proposed need scores corresponding to ordinal categories. The ANOVA approach may be directly applied to estimate the ICC for clustered ordinal data by imposing scores to ordered categories. Stiger et al. (1998) gave a detailed discussion on the assignment of integers when using ANOVA method to analyze ordinal data. Also, methods of scoring ordinal outcomes have been briefly introduced in section 1.4.2. In addition, one may assign weights to define the difference between ordinal distance. Cohen (1968) derived a weighted kappa statistic for ordinal data by using weights to describe the degree of disagreements among categories. Another method is to impose restrictions on odds ratios or probabilities to imply the ordinality. For example, one may derive the ICC estimators under ordinal regression models, e.g., the moment-based ICC estimator obtained from proportional odds models using GEE procedures.

The ICC estimators may be obtained by combining the above methods. For example, one has to assign both scores and weights to ordinal categories in order to obtain the weighted kappa statistic; to obtain the estimators from ordinal logistic regression models, it may be necessary to restrict odds ratios or probabilities and assign scores to categories. Additionally, there are close relationships among the three methods of imposing ordinality. For instance, Fleiss and Cohen (1973) established the equivalence of Cohen's weighted kappa with the quadratic weight and the two-way ANOVA ICC estimator.

In section 2.3.1, we introduce the ANOVA ICC estimation for clustered ordinal outcomes; in section 2.3.2, we briefly describe other estimation methods that have been used.

2.3.1 ANOVA method

Let Y_{ijl} denotes the ordinal score assigned to the *ijl*th observation. Consider a nested analysis of variance model given by $Y_{ijl} = \mu + \alpha_i + \gamma_{ij} + \varepsilon_{ijl}$. Random cluster effects, denoted by γ_{ij}, are assumed to be normally distributed with mean 0 and variance σ_c^2, i.e. $\gamma_{ij} \sim N(0, \sigma_c^2)$. We similarly assume the error terms $\varepsilon_{ijl} \sim N(0, \sigma_e^2)$. The ICC, ρ, may be interpreted as "the proportion of overall variation in response that can be accounted for by the between-cluster variation" (Donner and Klar, 2000, pp.8), i.e.,

$$\rho = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_e^2} \ .$$

The corresponding ANOVA table, which may be used to test the significance of the treatment effect, is shown in Table 2.1.

Table 2.1: Analysis of variance corresponding to a completely randomized design in which clusters are assigned to each of two intervention groups

	Degrees of freedom	Sum of squares (SS)	Mean square (MS)
Group	1	*SSG*	*MSG*
Clusters	$\sum_{i=1}^{2}(n_i - 1)$	*SSC*	*MSC*
Errors	$M - \sum_{i=1}^{2} n_i$	*SSE*	*MSE*
Total	$M - 1$	*SST*	

Here *MSC* and *MSE* are the between-cluster and within-cluster mean squares respectively, given by

$$MSC = \sum_{i=1}^{2}\sum_{j=1}^{n_i} m_{ij}(\overline{Y}_{ij} - \overline{Y}_i)^2 \,/(\sum_{i=1}^{2} n_i - 2)$$

and

$$MSE = \sum_{i=1}^{2}\sum_{j=1}^{n}\sum_{l=1}^{m_{ij}} (Y_{ijl} - \overline{Y}_{ij})^2 \,/(M - \sum_{i=1}^{2} n_i) \ .$$

Then the estimated ANOVA estimator ρ_A could be written as

$$\hat{\rho}_A = \frac{MSC - MSE}{MSC + (m_0 - 1)MSE} \qquad (2.1)$$

where

$$m_0 = \frac{M - \sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{m_{ij}^2}{M_i}}{\sum_{i=1}^{2} n_i - 2}.$$

2.3.2 Other methods

In addition to the ANOVA approach, the ICCs for cluster ordinal outcome data are often estimated by using model-based approaches. For instance, Lipsitz et al. (1994) proposed a moment-based approach to estimate ρ using generalized estimating equations (GEE) in proportional odds models. Let A_{ijl} be a diagonal matrix with the binary variances on the main diagonal, i.e.,

$$\hat{A}_{ijl} = Diag[\{\hat{P}_{ij1l}(1-\hat{P}_{ij1l}),....\hat{P}_{ij,K-1,l}(1-\hat{P}_{ij,K-1,l})\}]$$

and the residual matrix

$$\hat{e}_{ijl} = \hat{A}_{ijl}^{-\frac{1}{2}}[Y_{ijl} - \hat{P}_{ijl}].$$

Here $\hat{P}_{ijkl} = 1$ if $Y_{ijl} = k$ and 0 otherwise, and $\hat{P}_{ijl} = [\hat{P}_{ij1l}, \hat{P}_{ij2l},..., \hat{P}_{ij(K-1)l}]'$. Under a simple case of an exchangeable correlation structure, Lipsitz et al. (1994) derived

$$\hat{\rho}_{GEE} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{t>s}\hat{e}_{ijs}\hat{e}'_{ijt}}{[\sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{1}{2}m_{ij}(m_{ij}-1)]-3}$$

where $\hat{e}_{ijl}$ is estimated by substituting in $\hat{A}_{ijl}$ and $\hat{P}_{ijl}$ from a previous step of the Fisher scoring algorithm. We will further introduce the GEE approach in Chapter 4.

The ICC estimators may also be obtained by assuming a Dirichlet-multinomial model, e.g., a moment-based ICC estimator $\hat{\rho}_M$ by Lui et al. (1999). Consider *n* clusters are drawn from one single population and there are m_j observations in the *j*th cluster. Let the moment proportion estimator be

$$\hat{P}_k = \frac{\sum_{j=1}^{n}\sum_{l=1}^{m_j} 1(Y_{jl}, S_k)}{\sum_{j=1}^{n} m_j}$$

where $1(Y_{jl}, S_k) = 1$ if $Y_{jl} = S_k$, and $1(Y_{jl}, S_k) = 0$, otherwise. Then the stabilized moment ICC estimator is given by

$$\hat{\rho}_M = \frac{(n-1)[\sum_{k=1}^{K} S_k^2 \hat{P}_k - (\sum_{k=1}^{K} S_k \hat{P}_k)^2]}{MSC - (1-\varphi)[\sum_{k=1}^{K} S_k^2 \hat{P}_k - (\sum_{k=1}^{K} S_k \hat{P}_k)^2]} - 1$$

where φ is a shrinkage constant.

In addition to moment-based estimators, the ICC could also be estimated by using the MLE approach under dirichlet-multinomial models (Narayanan, 1991; Chuang and Cox, 1985; Paul et al. 2005). However, numerous authors (Tamura and Young, 1986; Tamura and Young, 1987; Yamamoto and Yanagimoto, 1992) noted that the MLE estimator generally underperforms the ANOVA and moment estimators for clustered binary outcome data with respect to the bias.

Additionally, one could derive the ICC estimator from the full likelihood function in a multivariate Plackett model (Molenberghs and Lesaffre, 1994). However, this approach requires sophisticated numerical procedures and it is difficult to implement in practice.

2.4 The ICC and the Measurement of Agreement

2.4.1 Introduction

The kappa statistic was developed to estimate interrater agreement for categorical outcomes, where interest focuses on the similarity among ratings obtained on the same subject. Scott (1955) proposed a chance-corrected measure of agreement between two raters by assuming that the marginal distribution of proportions over categories is equal for all raters. This index is often referred as Scott's π. Furthermore, Cohen (1960) extended Scott's π under the assumption of independent and potentially different marginal distribution of proportions for each rater. This statistic has come to be known as Cohen's kappa. In this study, we restrict our interests to Scott's π and Cohen's kappa. For other agreement measurements one could refer to the review by Banerjee et al. (1999).

Cohen (1968) generalized his kappa statistic to a weighted kappa by quantifying the severity of disagreement among ordinal categories. The most commonly used weights are "linear weights" and "quadratic weights" (Fleiss and Cohen, 1973). Furthermore, the weighted kappa statistic using quadratic weights is identical to the ICC estimator derived from a two-way ANOVA under the assumption that the subjects and the two rates are random samples from a universe of subjects and raters, respectively. As such, the relationship between kappa statistics and the ICC estimators has been built.

However, it is not appropriate to apply Cohen's weighted kappa to estimate the ICC in cluster randomization trials because there is rarely a natural order among cluster members. For instance, the *j*th subject from the *i*th cluster is a different individual in each cluster. Thus it is not possible to estimate separate marginal distributions for each rater. As an alternative, Scott's π assumes that the same marginal distribution of proportions for each rater. Therefore it is appropriate to use extensions of Scott's π to estimate the ICC for ordinal data in cluster randomization trials. The only exception would be the trials where cluster members can be ordered in some fashion so that the *j*th cluster member is the same in each cluster. For example, in the context of family randomization trials one could have the first subject is mother, the second one is dad and the third one is the first born child etc.

Note that both Scott's π and Cohen's kappa are derived from one single population, while the ICC estimates discussed here are from cluster randomization trials where there is one treatment group and one control group. As such one of challenges is to extend kappa statistics to two populations.

In next section, we propose a kappa-type ICC estimator for clustered ordinal data, denoted as $\hat{\rho}_\kappa$. In particular, we extend Scott's π statistic by using Abraira and De Vargas's (1999) approach. Generally there are three improvements in the new kappa-type ICC estimator compared with Scott's π: one is that weights are used to define the distance between ordinal categories; the second is that it suits well for variable cluster sizes by using pairwise agreement; and the third is that it allows treatment effects.

2.4.2 Kappa-type ICC Estimator

Scott's π was originally derived to measure agreement between two raters for multinomial outcomes. Let $p_{k\bullet}$ denotes the proportion of subjects placed in the *k*th category by the first rater, $p_{\bullet k}$ denotes the proportion of subjects placed in the *k*th category by the second rater, and p_k the proportion of the entire subjects falling in the *k*th category. Then, the kappa statistic proposed by Scott (1955) is defined as

$$\pi = \frac{p_o - p_e}{1 - p_e}. \quad (2.2)$$

Here

$$p_o = \sum_{k=1}^{K} p_{kk}$$

denotes the proportion of observed agreement and

$$p_e = \sum_{k=1}^{K} \left(\frac{p_{\bullet k} + p_{k\bullet}}{2} \right)^2$$

denotes the proportion of chance-expected agreement.

To extend Scott's π (1955) to clustered ordinal outcomes from trials where there is one treatment group and one control group, it is necessary to calculate P_o and P_e for each group separately. Let w_{gh} be the weight corresponding to the agreement between category g and h ($g, h = 1,2,...,K$), with the conditions:

$$0 \le w_{gh} < 1 \text{ for } g = h \text{ and } w_{gh} = 1 \text{ for } g \ne h .$$

For the *j*th cluster from the *i*th group, the number of weighted agreements is:

$$NA_{ij} = \frac{1}{2}\sum_{k=1}^{K} w_{kk} Y_{ijk}(Y_{ijk} - 1) + \sum_{g=1}^{K}\sum_{h>g}^{K} w_{gh} Y_{ijg} Y_{ijh} ,$$

and the number of possible pairs for the *ij*th cluster is:

$$\frac{1}{2} m_{ij}(m_{ij} - 1) .$$

Then the estimated proportion of weighted agreement for the *j*th cluster in the *i*th group is given by:

$$\frac{\frac{1}{2}\sum_{k=1}^{K} w_{kk} Y_{ijk}(Y_{ijk} - 1) + \sum_{g=1}^{K}\sum_{h>g}^{K} w_{gh} Y_{ijg} Y_{ijh}}{\frac{1}{2} m_{ij}(m_{ij} - 1)} .$$

Consequently the average observed weighted proportion of agreement for the *i*th group is given by

$$\hat{P}_{io} = \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{\frac{1}{2}\sum_{k=1}^{K} w_{kk} Y_{ijk}(Y_{ijk} - 1) + \sum_{g=1}^{K}\sum_{h>g}^{K} w_{gh} Y_{ijg} Y_{ijh}}{\frac{1}{2} m_{ij}(m_{ij} - 1)} \qquad (2.3).$$

Similarly the average expected proportion of pairwise agreement for the *i*th group is given by

$$\hat{P}_{ie} = \frac{\frac{1}{2}\sum_{k=1}^{K} w_{kk} Y_{ig}(Y_{ih} - 1) + \sum_{g=1}^{K}\sum_{h>g}^{K} w_{gh} Y_{ig} Y_{ih}}{\frac{1}{2} M_i (M_i - 1)}. \qquad (2.4)`$$

Thus the resulting kappa-type ICC estimator for the *i*th group is

$$\hat{\rho}_{i\kappa} = \frac{\hat{p}_{io} - \hat{p}_{ie}}{1 - \hat{p}_{ie}}.$$

To combine kappa-type ICC estimates of the two groups, Fleiss (1980, pp. 220-222) suggested an overall value

$$\hat{\rho}_{\kappa} = \frac{\sum_{i=1}^{2} w_i \hat{\rho}_{i\kappa}}{\sum_{i=1}^{2} w_i} = \frac{\sum_{i=1}^{2} \hat{P}_{io} - \sum_{i=1}^{2} \hat{P}_{ie}}{\sum_{i=1}^{2} (1 - \hat{P}_{ie})}. \qquad (2.5)$$

where the weight $w_i = 1 - \hat{P}_{ie}$.

Therefore the kappa-type ICC estimator for clustered ordinal outcomes is calculated with equation (2.5), using equation (2.3) and (2.4).

2.4.3 Connections with the ANOVA ICC Estimator

Assuming one single population, Fleiss and Cohen (1973) reported the identity between the ANOVA ICC estimator and the weighted kappa when there are only two observations in each cluster, using the quadratic weight

$$w_{gh} = 1 - \frac{(g-h)^2}{(K-1)^2}. \qquad (2.6)$$

Additionally, Fleiss (1981, pp. 226-pp.227) presented the asymptotical equivalence between the ANOVA ICC estimator and the kappa when outcomes have only two categories and cluster sizes are varying.

In previous sections we already derived the ICC estimator $\hat{\rho}_{\kappa}$ and $\hat{\rho}_{A}$ assuming the trials where there is one treatment group and one control group,. Here we explore the relationship between these two statistics.

Substituting w_{gh} into equation (2.3) and (2.4), $\hat{P}_{io}$ and $\hat{P}_{ie}$ may be rewritten as

$$\hat{P}_{io} = 1 - \sum_{j=1}^{n_i} \frac{2(\sum_{l=1}^{m_{ij}} Y_{ijl}^2 - m_{ij}\overline{Y}_{ij})}{n_i(m_{ij}-1)(K-1)^2}$$

and

$$\hat{P}_{ie} = 1 - \frac{2\sum_{j=1}^{n_i}\sum_{l=1}^{m_{ij}} Y_{ijl}^2 - 2M_i\overline{Y}_i^2}{(M_i-1)(K-1)^2}.$$

Thus the kappa-type ICC estimator in Equation (2.5) could be written as

$$\hat{\rho}_{\kappa} = 1 - \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{(\sum_{l=1}^{m_{ij}} Y_{ijl}^2 - m_{ij}\overline{Y}_{ij}^2)}{n_i(m_{ij}-1)}}{\sum_{i=1}^{2}\frac{\sum_{j=1}^{n_i}\sum_{l=1}^{m_{ij}} Y_{ijl}^2 - M_i\overline{Y}_i^2}{M_i-1}}. \quad (2.7)$$

In order to compare $\hat{\rho}_{\kappa}$ in (2.7) with $\hat{\rho}_{A}$ in (2.1), we restrict ourselves to a balanced cluster randomization trial (i.e., $m_{ij} = m$ and $n_i = n$). Thus $\hat{P}_{io}$ and $\hat{P}_{ie}$ reduce to

$$\widehat{P}_{io} = 1 - \frac{2(\sum_{j=1}^{n}\sum_{l=1}^{m} Y_{ijl}^2 - m\sum_{j=1}^{n}\overline{Y}_{ij}^2)}{(m-1)(K-1)^2 n}$$

and

$$\hat{P}_{ie} = 1 - \frac{2(\sum_{j=1}^{n}\sum_{l=1}^{m} Y_{ijl}^2 - mn\bar{Y}_i^{\cdot 2})}{(mn-1)(K-1)^2}$$

respectively. Therefore the kappa-type ICC estimator in equation (2.7) reduces to

$$\hat{\rho}_{\kappa} = \frac{MSC - MSE}{MSC + \frac{(m-1)MSE}{1-1/n}}. \qquad (2.8)$$

The ANOVA ICC estimator in equation (2.1) reduces to

$$\hat{\rho}_{A} = \frac{MSC - MSE}{MSC + (m-1)MSE}. \qquad (2.9)$$

Thus the two estimators are asymptotically equivalent as the number of clusters becomes large in a balanced trial. This result parallels Fleiss (1981, pp.226-227) and Fleiss and Cohen (1973)'s conclusions.

2.4.4 Properties

2.4.4.1 Reduction to Scott's π

Scott's π was originally derived to measure agreement between two raters and assumes that all disagreements among two different categories are equal. To reduce ρ_{κ} to Scott's π we have to extend the original Scott's π to allow the treatment effect first. We also need to limit our attention to the trials where there are two observations in a cluster (i.e., $m_{ij} = 2$) and the outcomes have two categories ($K = 2$) only. Thus the weight w_{gh} in $\hat{\rho}_{\kappa}$ is equal to 1 when $g = h$ and 0 otherwise.

For the *i*th group, the proportion of observed agreement in Scott's π is:

$$\hat{P}_{io} = \frac{1}{4n_i}\sum_{j=1}^{n_i}(Y_{ij1} - Y_{ij2})^2$$

and the proportion of chance-expected agreement is:

$$\hat{P}_{ie} = \frac{1}{4n_i^{\,2}}(Y_{i1}^2 + Y_{i2}^2).$$

Note that Y_{ij1} and Y_{ij2} here denotes the number of observations from the *ij*th cluster falling into the *k*th category, rather than the score assigned to the *ij*th observation.

Then Scott's $\hat{\pi}$ which allows the treatment effect is given by

$$\hat{\pi}_{overall} = \frac{\sum_{i=1}^{2}\hat{P}_{io} - \sum_{i=1}^{2}\hat{P}_{ie}}{\sum_{i=1}^{2}(1-\hat{P}_{ie})} = 1 - \frac{1 - \sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{1}{8n_i}(Y_{ij1} - Y_{ij2})^2}{1 - \sum_{i=1}^{2}\frac{1}{8n_i^2}(Y_{i1}^2 + Y_{i2}^2)}. \qquad (2.10)$$

The kappa-type ICC estimator $\hat{\rho}_{\kappa}$ in equation (2.5) reduces to

$$\hat{\rho}_{\kappa} = 1 - \frac{1 - \sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{1}{8n_i}(Y_{ij1} - Y_{ij2})^2}{1 - \sum_{i=1}^{2}\frac{1}{4n_i(2n_i-1)}(Y_{i1}^2 + Y_{i2}^2) + \sum_{i=1}^{2}\frac{1}{2(2n_i-1)}}. \quad (2.11)$$

Thus $\hat{\rho}_{\kappa}$ is asymptotically equal to $\pi_{overall}$ as the cluster number n_i becomes large.

When the number of clusters in each group is equal, i.e., $n_i = n$, the relationship between Scott's π and the kappa-type estimator could be shown more clearly as:

$$\hat{\pi}_{overall} = 1 - \frac{1 - \frac{1}{8n}\sum_{i=1}^{2}\sum_{j=1}^{n}(Y_{ij1} - Y_{ij2})^2}{1 - \frac{1}{8n^2}\sum_{i=1}^{2}(Y_{i1}^2 + Y_{i2}^2)} \qquad (2.12)$$

and

$$\hat{\rho}_{\kappa} = 1 - \frac{1 - \frac{1}{8n}\sum_{i=1}^{2}\sum_{j=1}^{n}(Y_{ij1} - Y_{ij2})^2}{1 - \frac{1}{4n\ (2n\ -1)}\sum_{i=1}^{2}(Y_{i1}^2 + Y_{i2}^2) + \frac{1}{2n\ -1}}. \qquad (2.13)$$

On the other hand, since Scott's π was originally derived from one single population, it may be of interest to derive the relationship of the two statistics by assuming one single population only (i.e., $i = 1$ and $n_i = n$). Thus Scott's $\hat{\pi}$ is given by

$$\hat{\pi} = 1 - \frac{1 - \frac{1}{8n}\sum_{j=1}^{2n}(Y_{j1} - Y_{j2})^2}{1 - \frac{1}{16n^2}(Y_1^2 + Y_2^2)}. \qquad (2.14)$$

Let $\hat{\rho}'_{\kappa}$ denotes the kappa-type ICC estimator from one single population, given by

$$\hat{\rho}'_{\kappa} = 1 - \frac{1 - \frac{1}{8n}\sum_{j=1}^{2n}(Y_{j1} - Y_{j2})^2}{1 - \frac{1}{4n(4n-1)}(Y_1^2 + Y_2^2) + \frac{1}{4n-1}}. \qquad (2.15)$$

The relationship between π and ρ'_{κ} parallels that from one single population.

In summary, we discussed the relationship between Scott's π and the kappa-type ICC estimator in this section. We first extended the original Scott's π to allow treatment effects. In order to simplify the formulas and show the relationship more clearly, we further assume equal number of clusters in each group. We also derived the relationship between the two statistics from one single population. We concluded that two statistics are asymptotically equivalent as the number of clusters becomes larger.

2.4.4.2 Minimum value

Fleiss (1981, pp. 225) derived a kappa statistic for binary data by applying the identity between intracluster correlation coefficients and kappa statistics. He further showed that his kappa statistic reaches the minimum value

$$\hat{\kappa} = -\frac{1}{\overline{m} - 1}.$$

when there is no variation across clusters in the proportion of positive ratings. Here $\overline{m} = \frac{M}{N}$. Similarly, we derive the minimum value of $\hat{\rho}_{\kappa}$ in this section.

When there is no variation among clusters in the proportions, under the alternative hypothesis that there is treatment effect, we have

$$\frac{Y_{ijk}}{m_{ij}} = \overline{P}_{ik}$$

for all *i,j,* with $\overline{P}_{ik}$ not equal to either 0 or 1. Let $\psi_i = \sum_{j=1}^{n_i} \sum_{l=1}^{m_{ij}} Y_{ijl}^2 - \overline{Y}^2 M_i$. Then $\hat{\rho}_{\kappa}$ in equation (2.5) reaches its minimum value:

$$\hat{\rho}_{\kappa(\min)} = 1 - \frac{\sum_{i=1}^{2} A_i \sum_{j=1}^{n_i} \frac{m_{ij}}{n_i (m_{ij} - 1) M_i}}{\sum_{i=1}^{2} A_i \frac{1}{(M_i - 1)}}. \quad (2.16)$$

To simplify the formula, we further assume there are equal number of clusters in each group and equal number of observations in each cluster. Thus the minimum value of $\hat{\rho}_{\kappa}$ in equation (2.16) reduces to

$$\hat{\rho}_{\kappa(\min)} = -\frac{1 - 1/n}{m - 1}. \quad (2.17)$$

Note that $\hat{\rho}_{\kappa(\min)}$ may be negative while the probability of obtaining a negative value becomes small as cluster sizes are large. Since negative ICC values are usually considered implausible in most application areas, it is common to set negative values to zero.

The minimum value in (2.16) or (2.17) is derived under the alternative hypothesis that there is treatment effect. However, under the null hypothesis of no treatment effect, i.e., $\frac{Y_{ijk}}{m_{ij}} = \overline{P}_{ik} = \overline{P}_k$ and $\overline{P}_k = \frac{Y_k}{M}$, the minimum value of $\hat{\rho}_{\kappa}$ is given by

$$\hat{\rho}_{\kappa(\min)} = 1 - \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{m_{ij}}{n_i(m_{ij}-1)}}{\sum_{i=1}^{2}\frac{M_i}{(M_i-1)}}. \quad (2.18)$$

When there are equal number of clusters in each group ($n_i = n$) and equal number of observation in each cluster ($m_{ij} = m$), it reduces to

$$\hat{\rho}_{\kappa(\min)} = -\frac{1-1/n}{m-1}.$$

It is equivalent to $\hat{\rho}_{\kappa(\min)}$ in equation (2.17).

2.4.4.3 Using midranks as scores

The calculation of kappa-type ICC estimator $\hat{\rho}_{\kappa}$ requires imposing scores to account for the order of the categories. We have briefly discussed methods of scoring in section 1.4.2. One of the scoring schemes, the equally spaced score, is frequently applied to obtain the ANOVA ICC estimator $\hat{\rho}_A$. As such we can use it in the kappa-type estimator so that $\hat{\rho}_{\kappa}$ could be related to $\hat{\rho}_A$. In addition, the score using midranks is one of methods that are commonly used in statistical procedures such as the Wilcoxon rank sum test. In particular, the Wilcoxon rank sum test and the Cochran-Armitage test are equivalent when midranks are assigned as scores. Therefore we also calculate $\hat{\rho}_{\kappa}$ by applying midranks as scores in this thesis so that different statistical methods may be unified in the next chapters.

Using equally spaced scores, for example, scores $1,2,...K$, the score for the *k*th category is $S_k = k$. Thus the quadratic weight in equation (2.5) is given by

$$w_{gh} = 1 - \frac{(g-h)^2}{(K-1)^2}. \quad (2.19)$$

By substituting it into $\hat{\rho}_{\kappa}$ in equation (2.5), we may obtain the kappa-type ICC estimator with scores $1,2,...K$.

Using midranks as scores, we have to calculate midranks for each group first. Under the alternative hypothesis that there are treatment effects, the midranks scoers from the two groups are different. Thus the midrank score for the *k*th category in the *i*th group is given by

$$S_{ik} = \sum_{g=1}^{k-1} Y_{ig} + (Y_{ik} + 1)/2 . \quad (2.20)$$

Consequently, the weight in the *i*th group is given by

$$w_{igh} = 1 - \frac{(S_{ig} - S_{ih})^2}{(\max(S_{i1}, S_{i2}, ...S_{iK}) - \min(S_{i1}, S_{i2}, ...S_{iK}))^2} . \quad (2.21)$$

In contrast, under the null hypothesis of no treatment effects, the midrank scores for the *k*th category from the two groups are identical, given by

$$S_{ik} = S_k = \sum_{g=1}^{k-1} Y_g + (Y_k + 1)/2 . \quad (2.22)$$

Consequently, the quadratic weight is given by

$$w_{igh} = w_{gh} = 1 - \frac{(S_g - S_h)^2}{(\max(S_1, S_2, ...S_K) - \min(S_1, S_2, ...S_K))^2} . \quad (2.23)$$

By substituting the weight w_{igh} in (2.22) and (2.23) into $\hat{\rho}_\kappa$, we may obtain the kappa-type ICC estimator with midrank scores under the alternative hypothesis and the null hypothesis correspondingly.

2.5 Summary

In section 2.3, we introduced methods which have been used to estimate the ICC for clustered ordinal data. In particular, we gave a detailed description of the ANOVA method.

In section 2.4, we proposed a kappa-type ICC estimator $\hat{\rho}_\kappa$ by extending Scott's by Abraira and Vargas's approach for clustered ordinal outcome data. Moreover, $\hat{\rho}_\kappa$ was

shown to be asymptotically equal to the ANOVA ICC estimator $\hat{\rho}_A$ as the number of clusters becomes large. We further discussed $\hat{\rho}_\kappa$'s properties, including its reduction to Scott's π, the minimum value, and options of imposed scores.

To summarize the ICC estimators discussed in this chapter, we list all ICC estimators Table 2.2. We will conduct simulation studies to evaluate $\hat{\rho}_A$ and $\hat{\rho}_\kappa$ and their relationships and properties in Chapter 6.

Table 2.2: Summary of the ICC estimators discussed in Chapter 2

Estimator	Method	General case	Special cases		Minimum values			
			$m_{ij} = m$ $n_i = n$	$m_{ij} = m = 2$ $n_i = n$ $k = 2$	Under H_A		Under H_0	
					General case	$m_{ij} = m$ $n_i = n$	General case	$m_{ij} = m$ $n_i = n$
$\hat{\rho}_A$	ANOVA ICC estimator	Equation (2.1) and (2.7)	Equation (2.9)					
$\hat{\rho}_\kappa$	kappa-type ICC estimator from two populations	Equation (2.5)	Equation (2.8)	Equation (2.13)	Equation (2.16)	Equation (2.17)	Equation (2.18)	Equation (2.17)
$\hat{\rho}'_\kappa$	kappa-type ICC estimator from one single population			Equation (2.15)				
$\hat{\pi}$	Scott's $\hat{\pi}$	Equation (2.2)		Equation (2.14)				
$\hat{\pi}_{overall}$	Scott's from two populations	Equation (2.10)		Equation (2.12)				

Chapter 3

3 Adjusted Cochran-Armitage Tests for Clustered Ordinal Outcomes

3.1 Introduction

In the previous chapter, we have presented methods for estimating the ICC for clustered ordinal outcome data. In the following chapters we discuss methods for analysis of clustered ordinal outcome data. We start with direct adjustment approaches which adapt simple corrections to the Cochran-Armitage test statistic for clustering effects.

The Cochran-Armitage trend test is a well-known approach for comparing binomial proportions among ordered groups. For independent ordinal outcome data, the Cochran-Armitage test statistic may equivalently be used to compare ordinal scores for two samples (Yates, 1948; Armitage, 1955). However, for clustered outcome data, the Cochran-Armitage trend test for comparing binary data can not be directly used to compare ordinal data.

She et al. (2010) extended the Cochran-Armitage test to genetic data from designs involving multistage cluster sampling. For each individual, they assigned the inverse of the product of the selection probabilities across all the stages of sampling as the weight. Then they adjusted all observed size in the Cochran-Armitage test statistic by the weights. However, its application to clustered ordinal outcomes was not discussed.

To extend the Cochran-Armitage test statistic for correlated ordinal data, Jung and Kang (2001) proposed a variance for the difference of scores between two groups that is obtained by standardizing the correlated scores. Although this approach takes into account the dependencies within clusters, the intraclass correlation coefficient (ICC) does not need to be specified.

Donner and Donald (1988) applied simple correction procedures to the Cochran-Armitage test to compare correlated binary outcomes on an ordinal cluster-level covariate. Their method, which utilizes an ICC for clustered binary data, offers such

advantages as simplicity and easy implementation. It does not necessarily require complicated computation and specified software. However, unlike the situation for independent outcome data, one cannot directly apply this adjusted test statistic to analyses of correlated ordinal data since the ICC for correlated binary outcome data is not equal to the ICC for correlated ordinal outcome data. Therefore a new ICC for correlated ordinal data must be used to obtain an adjusted version of the Cochran-Armitage trend test in this case.

In addition to Donner and Donald's approach, we extend the Cochran-Armitage trend test to clustered data using a weighted least squares approach. The Cochran-Armitage test was originally derived from a simple linear probability model by using the ordinary least squares approach (OLS) (Cochran, 1954; Armitage, 1955). However, the underlying assumptions of the OLS procedure are violated in cluster randomization trials where clustering induces a correlation among observations. In this case a more efficient estimator obtained by the weighted least square (WLS) approach may be used instead as an extension of the OLS procedure although the bias of estimator is unaffected by the choice of using OLS or WLS approach. Thus we adjust the Cochran-Armitage test to clustered outcome data by extending the OLS approach to a WLS approach.

In this chapter, we develop three simple adjustments to the regular Cochran-Armitage chi-square statistics for clustered binary data and clustered ordinal data respectively. The first one is Donner and Donald (1988)'s adjustment which is obtained by modifying the observed sample sizes of both the point estimate and its variance estimate in the test statistic; the second one is distinct in that it adjusts only the variance estimator in the statistic; the third one derives the statistic using a WLS approach. We list all six statistics in Table 3.1. The subscript 'CB' denotes clustered binary and 'CO' denotes clustered ordinal. In addition, the subscript '(1)' denotes the first adjustment method described above, '(2)' the second adjustment method, and '(3)' the third adjustment method.

The rest of the chapter is organized as follows. In section 3.2, we describe the Cochran-Armitage test for independent ordinal outcome data; in section 3.3, we present three

adjusted Cochran-Armitage tests for clustered binary outcome data; in section 3.4, we develop three adjusted Cochran-Armitage trend tests for clustered ordinal outcome data.

3.2 Cochran-Armitage Test for Independent outcomes

Suppose there are G ordered groups consisting of subjects having binary outcomes. Let S_i be a score variable which is associated with the ith group, $i=1,2,\ldots,G$. Let A_i denotes the number of successes in the ith group and M_i denotes the total number of observations in the ith group. Then the proportion of successes in group i is given by

$$\hat{P}_i = A_i \big/ M_i .$$

Table 3.1: Summary of the Cochran-Armitage trend tests in Chapter 3

Test statistic	Method	Approach	Formula	Outcome data
χ^2	Cochran-Armitage test	Ordinary least squares	Equation (3.1)	Independent data
$\chi^2_{CB-(1)}$	Donner and Donald's test	Adjusting point estimator and its variance estimator	Equation (3.3)	Clustered binary data
$\chi^2_{CB-(2)}$	An Alternative to Donner and Donald's Test	Adjusting the variance estimate only	Equation (3.5)	
χ^2_{CB-WLS}	Weighted-Least-Square Cochran-Armitage Test	Weighted least squares	Equation (3.7)	
$\chi^2_{CO-(1)}$	Donner and Donald's test	Adjusting point estimator and its variance estimator	Equation (3.9)	Clustered ordinal data
$\chi^2_{CO-(2)}$	An Alternative to Donner and Donald's Test	Adjusting the variance estimate only	Equation (3.11)	
χ^2_{CO-WLS}	Weighted-Least-Square Cochran-Armitage Test	Weighted least square	Equation (3.13)	

Let S_i be the score variable associated with the *i*th group. Then the linear probability model Cochran and Armitage used to evaluate the trend in the proportion of success $\hat{P}_i$ with S_i is

$$E(\hat{P}_i) = \alpha + \beta S_i$$

where α and β are the intercept and slope parameters. Since our objective is to test the null hypothesis of no trend, i.e., $H_0 : \beta = 0$, we omit inferences about α and only focus on β in this research.

In the case of independent outcomes, the ordinary least squares estimator of β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{G} M_i(\hat{P}_i - \overline{P})(S_i - \overline{S})}{\sum_{i=1}^{G} M_i(S_i - \overline{S})^2}.$$

Here $\overline{S}$ denotes the mean values of *S* given by

$$\overline{S} = \frac{\sum_{i=1}^{G} M_i S_i}{\sum_{i=1}^{G} M_i},$$

and $\overline{P}$ denotes the overall proportion of success given by

$$\overline{P} = \frac{\sum_{i=1}^{G} M_i \hat{P}_i}{\sum_{i=1}^{G} M_i}.$$

Under the null hypothesis $H_0 : \beta = 0$, the corresponding least squares variance estimator of $\hat{\beta}$ is

$$\widehat{\text{var}}(\hat{\beta}) = \frac{\bar{P}(1-\bar{P})}{\sum_{i=1}^{G} M_i (S_i - \bar{S})^2}.$$

Thus a one degree of freedom Cochran-Armitage trend test statistic is given by

$$\chi^2 = \frac{\hat{\beta}^2}{\widehat{\text{var}}(\hat{\beta})} = \hat{\beta}^2 \sum_{i=1}^{G} M_i (S_i - \bar{S})^2 / \bar{P}(1-\bar{P}) \ . \qquad (3.1)$$

It may also be derived from a simple linear model regressing a binary outcome on an ordinal covariate (Bland, 2000; pp243):

$$Y_i = \alpha + \beta S_i \, .$$

Here Y_i=1 if the *i*th observation is a "success" and 0 otherwise. Moreover, it is also equivalent to the score statistic obtained from logistic regression analyses with a single covariate (Cox, 1958).

Additionally, the Cochran-Armitage test is related to a variety of non-parametric and model-based methods. For example, it is equivalent to the Wilcoxon rank sum test when the scores are set equal to the midranks. It is also equivalent to the Mantel extension test (Mantel, 1963), explaining why it is frequently called the Cochran-Armitage-Mantel trend test. We will further discuss relationships between the Cochran-Armitage test and other test statistics in Chapter 5.

3.3 Adjusted Cochran-Armitage test for clustered binary outcome data

3.3.1 Donner and Donald's Test

We assume that it is of interest to compare *G* groups consisting of observed binary outcomes. Suppose that n_i clusters (*i=1,2,…, G*) are randomized to the *i*th group. Let m_{ij} denote the size of the *ij*th (*j=1,2,…,* n_i) cluster and $a_{ij} = \sum_{l=1}^{m_{ij}} Y_{ijl}$ the number of successes in the *ij*th cluster. Denote the total number of individuals in the *i*th group by $M_i = \sum_{j=1}^{n_i} m_{ij}$ and the corresponding total number of successes by $A_i = \sum_{j=1}^{n_i} a_{ij}$. Then $\hat{P}_i = A_i / M_i$ denotes the proportion of successes in the *i*th group. The resulting data layout is

presented in Table 3.2. This table was originally developed by Donner and Banting (1989).

Table 3.2: Data lay-out for adjusted cochran-armitage test for clustered binary outcomes

Group	Number of Clusters	Number of Observations	Number of Observations with Characteristic	Proportion of Observations with Characteristic
1	n_1	$M_1 = \sum_{j=1}^{n_i} m_{1j}$	$A_1 = \sum_{j=1}^{n_i} a_{1j}$	$\hat{P}_1 = A_1 / M_1$
2	n_2	$M_2 = \sum_{j=1}^{n_2} m_{2j}$	$A_2 = \sum_{j=1}^{n_2} a_{2j}$	$\hat{P}_2 = A_2 / M2$
⋮	⋮	⋮	⋮	⋮
G	n_G	$M_G = \sum_{j=1}^{n_G} m_{Gj}$	$A_G = \sum_{j=1}^{n_G} a_{Gj}$	$\hat{P}_G = A_G / M_G$
Total	N	$\sum_{i=1}^{G} M_i = M$	$\sum_{i=1}^{G} A_i = A$	$\overline{P} = A / M$

The linear probability model used to test the trend for clustered binary outcome data is written as

$$E(\hat{P}_i) = \alpha_C + \beta_C S_i . \quad (3.2)$$

To adjust the test statistic χ^2 to clustered binary data, Donner and Donald (1988) replaced the observed sample size M_i by $M_i / \overline{C}_{i(B)}$. Here $\overline{C}_{i(B)}$ denotes the design effect, also referred to as "variance inflation factor" indicating the variance of the success rate in each group increases as a result of clustering, given by

$$\overline{C}_{i(B)} = \frac{\sum_{j=1}^{n_i} m_{ij}[1 + (m_{ij} - 1)\hat{\rho}_B]}{\sum_{j=1}^{n_i} m_{ij}} = 1 + (\overline{m}_{Ai} - 1)\hat{\rho}_B .$$

Here $\hat{\rho}_B$ is an estimator of the ICC for clustered binary data and $\overline{m}_{Ai} = \sum_{j=1}^{n_i} m_{ij}^2 / \sum_{j=1}^{n_i} m_{ij}$.

Thus the slope parameter estimator of β_C is given by

$$\hat{\beta}_{CB-(1)} = \frac{\sum_{i=1}^{G} \frac{M_i}{\overline{C}_{i(B)}} (\hat{P}_i - \overline{P})(S_i - \overline{S})}{\sum_{i=1}^{G} \frac{M_i}{\overline{C}_{i(B)}} (S_i - \overline{S})^2}.$$

Under the null hypothesis $H_0 : \beta_C = 0$, the corresponding variance estimator is

$$\widehat{\text{var}}(\hat{\beta}_{CB-(1)}) = \frac{\overline{P}(1-\overline{P})}{\sum_{i=1}^{G} \frac{M_i}{\overline{C}_{i(B)}} (S_i - \overline{S})^2}.$$

Consequently the Cochran-Armitage trend test for clustered binary outcome data is given by

$$\chi_{CB-(1)}{}^2 = \frac{\hat{\beta}_{CB-(1)}{}^2}{\widehat{\text{var}}(\hat{\beta}_{CB-(1)})} = \frac{[\sum_{i=1}^{G} (S_i - \overline{S})(\hat{P}_i - \overline{P}) \frac{M_i}{\overline{C}_{i(B)}}]^2}{\overline{P}(1-\overline{P}) \sum_{i=1}^{G} (S_i - \overline{S})^2 \frac{M_i}{\overline{C}_{i(B)}}}. \qquad (3.3)$$

To estimate the unknown ICC parameter $\hat{\rho}_B$ for clustered binary data, Donner and Donald (1988) considered the use of the ANOVA approach. Let $M = \sum_{i=1}^{G} \sum_{j=1}^{n_i} m_{ij}$ denotes the total individuals in the study, $N = \sum_{i=1}^{G} n_i$ denotes the total clusters, and

$$\hat{P}_{ij} = a_{ij} / m_{ij}$$

denotes the proportion of successes in the *ij*th cluster. Then the mean square errors between and within clusters in the case of binary outcome data are given, respectively, by:

$$MSC = \frac{\sum_{i=1}^{G}\sum_{j=1}^{n_i} m_{ij}(\hat{P}_{ij} - \hat{P}_i)^2}{N - G}$$

and

$$MSC = \frac{\sum_{i=1}^{G}\sum_{j=1}^{n_i} m_{ij}\hat{P}_{ij}(1 - \hat{P}_{ij})}{M - N}.$$

Then the ANOVA ICC estimator for clustered binary outcomes is given by

$$\hat{\rho}_B = \frac{MSC - MSW}{MSC + (m_0 - 1)MSW}$$

where $m_0 = (M - \sum_{i=1}^{G}\sum_{j=1}^{n_i} m_{ij}^2 / M_i)/(G-2)$.

In general, the ANOVA estimator and the moment estimators are two simple approaches which do not involve sophisticated computation. Thus Donner and Donald (1988) compared these two statistics and simulation results showed that the ANOVA estimator tended to be more accurate than the moment-based estimator in terms of mean square error. Hence the ANOVA ICC estimator $\hat{\rho}_B$ was considered by Donner and Donald (1988) to use in their adjusted Cochran-Armitage test.

Note that there is a typographical error in Donner and Donald (1988)'s paper. They denoted M_i as the number of clusters randomly assigned to the *i*th group. Actually, it should be correctly referred to the number of observations in the *i*th group. This typographical error was corrected by Donner and Banting (1988).

When individuals in a cluster are statistically independent of each other, i.e., $\rho_B = 0$, Donner and Donald (1988)'s adjusted test statistic $\chi^2_{CB-(1)}$ reduces to the regular Cochran-Armitage test statistic χ^2. Additionally, when there are only two groups, e.g., *G=2*, $\chi^2_{CB-(1)}$ reduces to

$$\chi^2_{CB-(1)} = \frac{M_1 M_2}{(M_1 + M_2)} \left(\frac{M_1}{C_1} + \frac{M_2}{C_2} \right) \frac{(\hat{P}_1 - \hat{P}_2)^2}{\overline{P}(1-\overline{P})}.$$

It is identical to the adjusted Pearson chi-square test proposed by Donner and Donald (1988).

Moreover, when cluster sizes are constant, $\chi^2_{CB-(1)}$ reduces to

$$\chi^2_{CB-(1)} = \frac{[\sum_{i=1}^{G} (S_i - \overline{S})(\hat{P}_i - \overline{P}) M_i]^2}{\overline{P}(1-\overline{P}) \sum_{i=1}^{G} (S_i - \overline{S})^2 M_i [1 + (m-1)\hat{\rho}_B]} = \frac{\chi^2}{1 + (m-1)\hat{\rho}_B} \quad (3.4)$$

It is simply the division of the regular Cochran-Armitage test statistic by $1 + (m-1)\hat{\rho}_B$.

3.3.2 An Alternative to Donner and Donald's Test

There are a variety of ways to adjust a test statistic for a clustering effect. For example, adjustments of the C-A test to clustered binary outcomes include Donner and Donald (1988), Rao and Scott (1992), Fung et al. (1994), Jung and Kang (2001), Stefanescu and Turnbull (2003) and She et al. (2010).

In addition, there are two general ways to adjust the test statistic which is obtained by dividing the point estimator by its variance estimator. One is to adjust the variance estimate only, and the other is to adjust both the point estimator and its variance estimate. Related discussions include Scott and Holt (1982), Donner and Klar (2001, pp.90-91) and Zou (2002, pp.29-32). In particular, Scott and Holt (1982) discussed these two adjustments for clustered continuous outcome data in the context of linear regression. They compared the OLS parameter estimate and its variance estimate with the weighted least-squares (WLS) parameter estimate and the corresponding variance estimate. They reported that the OLS variance estimator is seriously biased and then affects the hypothesis testing procedures. Thus it should be substituted by the WLS variance estimator in order to guarantee validity. However, the OLS estimators of regression coefficients remain unbiased and are fairly efficient when the ICC is small and cluster

sizes are large. Thus both the OLS and WLS estimators of regression coefficients may be used in test procedures based on them.

The C-A trend test statistic is derived by dividing the parameter estimator by its variance estimator in the context of linear trend model. Donner and Donald (1988) modified the C-A test statistic by adjusting observed sample sizes in both the numerator and denominator of the standard C-A statistic for clustering effect. Hence both the point estimator of β_C and its variance estimator are adjusted by the variance inflation factor $\overline{C}_{i(B)}$. In this section, we propose an approach which is also based on a simple adjustment of the standard C-A test. However, unlike Donner and Donald's (1988) method, this approach only adjusts the variance estimator while not adjusting the point estimator of β_C.

Thus the slope estimate in the linear probability trend model (3.2) is given by

$$\hat{\beta}_{CB-(2)} = \frac{\sum_{i=1}^{G} M_i(\hat{P}_i - \overline{P})(S_i - \overline{S})}{\sum_{i=1}^{G} M_i(S_i - \overline{S})^2}.$$

Under H_0, its corresponding adjusted variance estimator is

$$\widehat{\mathrm{var}}(\hat{\beta}_{CB-(2)}) = \frac{\overline{P}(1-\overline{P})\sum_{i=1}^{G} M_i(S_i - \overline{S})^2 \overline{C}_{i(B)}}{[\sum_{i=1}^{G} M_i(S_i - \overline{S})^2]^2}.$$

Therefore the adjusted trend test statistic is given by

$$\chi_{CB-(2)}{}^2 = \frac{[\sum_{i=1}^{G}(S_i - \overline{S})(\hat{P}_i - \overline{P})M_i]^2}{\overline{P}(1-\overline{P})\sum_{i=1}^{G}(S_i - \overline{S})^2 M_i \overline{C}_{i(B)}}. \qquad (3.5)$$

When the cluster sizes are equal, $\chi^2_{CB-(2)}$ reduces to

$$\chi_{CB-(2)}{}^2 = \frac{\chi^2}{1+(m-1)\hat{\rho}_B}.$$

It is identical to $\chi^2_{CB-(1)}$ in equation (3.4). We will further compare $\chi^2_{CB-(1)}$ and $\chi^2_{CB-(2)}$ in case of varying cluster sizes by simulated data in Chapter 6. Their performance will be evaluated in terms of simulated Type I error and power.

In addition to the two adjustments presented here, Stefanescu and Turnbull (2003) generalized the C-A test to assess the trend among clusters. To relate their test statistic to statistics proposed here, we need to assume the sizes of the clusters are equal in each group. Thus their statistic testing for trend among clusters could be linked to Donner and Donald's (1988) statistic testing for trend among groups.

Stefanescu and Turnbull (2003)'s test statistic is given by

$$\chi^2_{ST} = \frac{[\sum_{i=1}^{G}(S_i - \bar{S})(\hat{P}_i - \bar{P})M_i]^2}{\bar{P}(1-\bar{P})\sum_{i=1}^{G}(S_i - \bar{S})^2 M_i \bar{C}_{i(B)}}. \qquad (3.6)$$

It is identical to $\chi^2_{CB-(2)}$ in equation (3.5).

3.3.3 Weighted Least Squares Cochran-Armitage Test

The Cochran-Armitage trend test was originally derived from a linear probability model by using the ordinary least squares (OLS) approach (Cochran, 1954; Armitage, 1955). However, when the OLS approach is applied to cluster randomization trials the variance estimates may be seriously biased and therefore inference procedures based on these estimates can be misleading. As a result, the WLS approach is often used as an extension of OLS to account for clustering effects.

It is straightforward to understand the underlying nature of the use of weighted least squares approach in cluster randomization trials. For instance, a proper weight is given to a cluster according to its variance so that more variable observations in a cluster contribute less to data information than do less variable observations in a cluster. Hence, we consider this approach in extending the Cochran-Armitage test to clustered outcome data in this section.

More over, one appealing feature of the WLS approach is that it does not require complex computation and specialized software. It also has close connections with more sophisticated methods. For example, maximum likelihood estimation (MLE) algorithms (e.g., Fisher scoring algorithm) often consist of iterative use of WLS. Also, Agresti et al (1991) reported that when the marginal models for categorical outcomes hold, MLE and WLS estimates are asymptotically equivalent with large cell expected frequencies. Additionally, Miller et al. (1993) illustrated that the WLS estimate is the first iteration result of the GEE procedure.

In this section we derive the adjusted C-A test for clustered binary data using the weighted least squares (WLS) approach.

Under the null hypothesis $H_0 : P_1 = P_2 = ... = P_G = P$, Y_{ijl} has a variance of $\sigma^2 = \overline{P}(1-\overline{P})$. Let V_{ij} represent the variance matrix for a single cluster given by

$$V_{ij} = \sigma^2\{(1-\rho_B)I + \rho_B J\}$$

where I denotes a $m_{ij} \times m_{ij}$ identity matrix and J the $m_{ij} \times m_{ij}$ matrix all of whose elements are 1. Let V be a block-diagonal variance matrix with non-zero $m_{ij} \times m_{ij}$ blocks V_{ij}. We denote W as the $M \times M$ weight matrix for the WLS approach, where

$$W = V^{-1}.$$

Then still consider the model used to evaluate the trend for clustered binary data in (3.2). The WLS estimator of β_C is given by

$$\hat{\beta}_{CB-WLS} = \frac{\sum_{i=1}^{G}(S_i - \tilde{S})(\hat{P}_i - \tilde{P})\sum_{j=1}^{n_i}\frac{m_{ij}}{[1+(m_{ij}-1)\hat{\rho}_B]}}{\sum_{i=1}^{G}(S_i - \tilde{S})^2\sum_{j=1}^{n_i}\frac{m_{ij}}{[1+(m_{ij}-1)\hat{\rho}_B]}}.$$

Here

$$\tilde{S} = \frac{\sum_{i=1}^{G} S_i \sum_{j=1}^{n} \frac{m_{ij}}{[1+(m_{ij}-1)\hat{\rho}_B]}}{\sum_{i=1}^{G}\sum_{j=1}^{n_i} \frac{m_{ij}}{[1+(m_{ij}-1)\hat{\rho}_B]}} \text{ and } \tilde{P} = \frac{\sum_{i=1}^{G} \hat{P}_i \sum_{j=1}^{n} \frac{m_{ij}}{[1+(m_{ij}-1)\hat{\rho}_B]}}{\sum_{i=1}^{G}\sum_{j=1}^{n_i} \frac{m_{ij}}{[1+(m_{ij}-1)\hat{\rho}_B]}}.$$

The derivation of $\hat{\beta}_{CB-WLS}$ is provided in Appendix A. Under H_0, the corresponding estimated variance is

$$\widehat{\text{var}}(\hat{\beta}_{CB-WLS}) = \frac{\tilde{P}(1-\tilde{P})}{\sum_{i=1}^{G}(S_i-\tilde{S})^2 \sum_{j=1}^{n_i} \frac{m_{ij}}{1+(m_{ij}-1)\hat{\rho}_B}}.$$

Consequently the chi-square trend test statistic derived from the WLS approach is given by

$$\chi_{CB-WLS}^{\ 2} = \frac{\hat{\beta}_{CB-WLS}^{\ 2}}{\widehat{\text{var}}(\hat{\beta}_{CB-WLS})} = \frac{[\sum_{i=1}^{G}(S_i-\tilde{S})(\hat{P}_i-\tilde{P})\sum_{j=1}^{n_i}\frac{m_{ij}}{1+(m_{ij}-1)\hat{\rho}_B}]^2}{\tilde{P}(1-\tilde{P})\sum_{i=1}^{G}(S_i-\tilde{S})^2\sum_{j=1}^{n_i}\frac{m_{ij}}{1+(m_{ij}-1)\hat{\rho}_B}}. \qquad (3.7)$$

As an alternative to using the weight $W = V^{-1}$, one may use the observed cluster sizes m_{ij} or the "effective sample size" $\frac{m_{ij}}{\overline{C}_{i(B)}}$ as the weight. However, the most efficient WLS estimator of β uses $W = V^{-1}$, which we adopted here.

In the special case of $\rho_B = 0$, the WLS trend test reduces to the regular Cochran-Armitage test. When the cluster sizes are constant, i.e., $m_{ij} = m$, $\tilde{P} = \overline{P}$, and $\tilde{S} = \overline{S}$, the WLS trend test statistic reduces to

$$\chi_{CB-WLS}^{\ 2} = \frac{[\sum_{i=1}^{G}(S_i-\overline{S})(\hat{P}_i-\overline{P})\frac{M_i}{1+(m-1)\hat{\rho}_B}]^2}{\overline{P}(1-\overline{P})\sum_{i=1}^{G}(S_i-\overline{S})^2\frac{M_i}{1+(m-1)\hat{\rho}_B}} = \frac{\chi^2}{1+(m-1)\hat{\rho}_B}.$$

Furthermore, when there are only two groups and the cluster sizes are constant as well, the statistic χ^2_{CB-WLS} reduces to

$$\chi^2_{CB-WLS} = \frac{M_1 M_2}{(M_1 + M_2)}\left(\frac{M_1}{C_1} + \frac{M_2}{C_2}\right)\frac{(\hat{P}_1 - \hat{P}_2)^2}{\bar{P}(1-\bar{P})}.$$

It is identical to Donner and Donald (1988)'s adjusted Pearson Chi-square statistic.

In addition, we derive the relationship between the WLS C-A test statistic and the score test statistic derived from a binary logistic regression by using the GEE. When there are only two groups, χ^2_{CB-WLS} reduces to

$$\chi^2_{CB-WLS} = \frac{(\hat{P}_1 - \hat{P}_2)^2}{\tilde{P}(1-\tilde{P})\sum_{i=1}^{2}\frac{1}{\sum_{j=1}^{n_i}\frac{m_{ij}}{C_{ij}}}}.$$

The score test statistic, which is derived from a binary logistic regression using the GEE and assuming an exchangeable working correlation matrix, is given by

$$\chi^2_{GEE(score)} = \frac{(\tilde{P}_1 - \tilde{P}_2)^2}{\tilde{P}(1-\tilde{P})\sum_{i=1}^{2}\frac{1}{\sum_{j=1}^{n_i}\frac{m_{ij}}{C_{ij}}}}$$

where $\tilde{P}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}/C_{ij}}{\sum_{j=1}^{n_i} m_{ij}/C_{ij}}$.

3.4 Adjusted Cochran-Armitage Test for Clustered Ordinal Outcomes

We presented three adjusted C-A trend tests for clustered binary data in the previous section. In this section, we correspondingly extend these three methods to clustered ordinal outcome data.

3.4.1 Extension of Donner and Donald's Test

We assume that it is of interest to compare two groups consisting of ordinal outcomes. Suppose n_i clusters are randomly assigned to the *i*th group, *i=1* or *2,* where there are m_{ij} observations in the *ij*th cluster. Each observation may have an outcome in any of *K* categories. Let Y_{ijk} be the number of observations falling into the *k*th category from the *ij*th cluster, $j = 1,...,n_i$ and $k = 1,...,K$. Let A_{ijk} denote the number of observations falling into the *k*th category from the *ij*th cluster that have the characteristic. Then

$$A_k = \sum_{i=1}^{2} \sum_{j=1}^{n_i} A_{ijk}$$

and

$$Y_k = \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ijk} .$$

Hence the proportion of successes from the *k*th category is given by

$$\hat{P}_k = \frac{A_k}{Y_k} .$$

The resulting data layout is presented in Table 3.3 and 3.4.

Let $\overline{P}'$ denote the overall proportion of successes given by

Table 3.3: Data lay-out for adjusted Cochran-Armitage test for clustered ordinal outcomes

Group	Cluster	Number of observations	Number of observations in the *kth* category	Number of observations with characteristic in the *kth* category	Proportion in the *kth* category
1	1	m_{11}	Y_{11k}	A_{11k}	$\hat{P}_{11k} = A_{11k} / Y_{11k}$
	2	m_{12}	Y_{12k}	A_{12k}	$\hat{P}_{12k} = A_{12k} / Y_{12k}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n_1	m_{1n_1}	Y_{1n_1k}	A_{1n_1k}	$\hat{P}_{1n_1k} = A_{1n_1k} / Y_{1n_1k}$
2	1	m_{21}	Y_{21k}	A_{21k}	$\hat{P}_{21k} = A_{21k} / Y_{21k}$
	2	m_{22}	Y_{22k}	A_{22k}	$\hat{P}_{22k} = A_{22k} / Y_{22k}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n_2	m_{2n2}	Y_{2n_2k}	A_{2n_2k}	$\hat{P}_{2n_2k} = A_{2n_2k} / Y_{2n_2k}$
Total	N	$M = \sum_{i=1}^{2}\sum_{j=1}^{n_i} m_{ij}$	$Y_k = \sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ijk}$	$A_k = \sum_{i=1}^{2}\sum_{j=1}^{n_i} A_{ijk}$	$\hat{P}_k = A_k / Y_k$

Table 3.4: Data lay-out for clustered ordinal outcomes in the *ij*th cluster

Outcome	Number of observations with characteristic	Number of observations without characteristic	Total
$k = 1$	A_{ij1}	$Y_{ij1} - A_{ij1}$	Y_{ij1}
$k = 2$	A_{ij2}	$Y_{ij2} - A_{ij2}$	Y_{ij2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k = K$	A_{ijK}	$Y_{ijK} - A_{ijK}$	Y_{ijK}

$$\overline{P}' = \frac{\sum_{k=1}^{K} A_k}{\sum_{k=1}^{K} Y_k}.$$

Let S_k denote the score associated with the *k*th category and

$$\overline{S}' = \frac{\sum_{k=1}^{K} S_k Y_k}{\sum_{k=1}^{K} Y_k}.$$

Here we use the superscript ' to distinguish $\overline{P}'$ and $\overline{S}'$ with $\overline{P}$ and $\overline{S}$ calculated for clustered binary outcomes in section 3.3.

The model used to test trend among ordered categories, or the equivalence between the ordinal outcomes from the two groups, is given by

$$E(\hat{P}_k) = \alpha_C + \beta_C S_k \qquad (3.8)$$

with the null hypothesis $H_0 : \beta_C = 0$. Let $\hat{\rho}_O$ denote the ICC estimator for clustered ordinal outcomes and then the variance inflation factor in the *i*th group may be written as

$$\overline{C}_{i(O)} = \frac{\sum_{j=1}^{n_i} m_{ij}[1 + (m_{ij} - 1)\hat{\rho}_O]}{\sum_{j=1}^{n_i} m_{ij}}.$$

Thus the adjusted slope estimator in model (3.8) is given by

$$\hat{\beta}_{CO-(1)} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k - \overline{S}')(\hat{P}_k - \overline{P}')\frac{Y_{ijk}}{\overline{C}_{i(O)}}}{\sum_{i=1}^{2}\sum_{j=1}^{n}\sum_{k=1}^{K}(S_k - \overline{S}')^2\frac{Y_{ijk}}{\overline{C}_{i(O)}}}.$$

Under H_0, the corresponding variance estimator is

$$\widehat{\text{var}}(\hat{\beta}_{CO-(1)}) = \frac{\bar{P}'(1-\bar{P}')}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k-\bar{S}')^2\frac{Y_{ijk}}{\bar{C}_{i(O)}}}.$$

Consequently, the adjusted C-A test for clustered ordinal outcome data is given by

$$\chi^2_{CO-(1)} = \frac{\hat{\beta}^2_{CO-(1)}}{\widehat{\text{var}}(\hat{\beta}_{CO-(1)})} = \frac{[\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k-\bar{S}')(\hat{P}_k-\bar{P}')\frac{Y_{ijk}}{\bar{C}_{i(O)}}]^2}{\bar{P}'(1-\bar{P}')\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k-\bar{S}')^2\frac{Y_{ijk}}{\bar{C}_{i(O)}}} \qquad (3.9)$$

It may be easily shown that the statistic $\chi^2_{CO-(1)}$ follows a chi-square distribution with one degree of freedom under H_0.

Note that the linear trend model for clustered ordinal outcomes in (3.8) is same as the one for clustered binary outcomes in (3.2), except that the subscript '*k*' in (3.8) denotes the *k*th category in which the ordinal outcomes fall while the subscript '*i*' in the model (3.2) denotes the *i*th group in which cluster are randomized. As such the null hypotheses in these two linear trend models are identical, given by $H_0 : \beta_C = 0$. We presume this identity between the two models and their null hypotheses since the same hypothesis is used for independent binary and ordinal outcomes. The appropriateness of this presumption may need further considerations while it assures the most convenient way to handle the question at hand.

However, the underlying meaning of the models and the corresponding null hypothesis in each model depend on what outcomes the ICC is adopted to estimate. If we adopt ρ_B to analyze clustered binary outcomes, the models in (3.2) and (3.8) would be used to test the trend among *G* (*i=1,2,…,G*) groups or trend among *K* (*k=1,2,…,K*) categories. The corresponding null hypothesis is that there is no trend among *G* (*i=1,2,…,G*) groups or there is no trend among *K* (*k=1,2,…,K*) categories. In contrast, if we substitute in ρ_O to analyze clustered ordinal outcomes, the models in (3.2) and (3.8) would test the equality between the ordinal outcomes between the two groups. Thus the corresponding null

hypothesis is interpreted as there is no difference between ordinal outcomes from two samples.

Methods of estimating the ICC ρ_O have been discussed previously in Chapter 2. In this research we restrict our attention to the ANOVA ICC estimator and the kappa-type estimator presented in section 2.3 and 2.4. We will then evaluate the performance of the adjusted Cochran-Armitage test with the use of these two ICC estimators in simulation studies.

In the special case of $\rho_O = 0$, the statistic $\chi^2_{CO-(1)}$ reduces to the regular Cochran-Armitage test statistic. When cluster sizes are equal, i.e., $m_{ij} = m$, the adjusted trend test statistic reduces to

$$\chi_{CO-(1)}{}^2 = \frac{[\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k - \bar{S}')(\hat{P}_k - \bar{P}')Y_{ijk}]^2}{\bar{P}'(1-\bar{P}')\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k - \bar{S}')^2 Y_{ijk}[1+(m-1)\hat{\rho}_O]} = \frac{\chi^2}{1+(m-1)\hat{\rho}_O}. \qquad (3.10)$$

Additionally, when the outcomes have only two categories and cluster sizes are constant as well, $\chi^2_{CO-(1)}$ reduces to Donner and Donald (1988)'s adjusted Pearson test statistic.

3.4.2 Extension of An Alternative to Donner and Donald's Test

We now extend the adjusted C-A statistic $\chi^2_{CB-(2)}$ to clustered ordinal data. Consider the linear trend model in (3.8). Then the slope estimator without adjusting is given by

$$\hat{\beta}_{CO-(2)} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k - \bar{S}')(\hat{P}_k - \bar{P}')Y_{ijk}}{\sum_{i=1}^{2}\sum_{j=1}^{n}\sum_{k=1}^{K}(S_k - \bar{S}')^2 Y_{ijk}}.$$

Under H_0, the corresponding variance estimator is

$$\widehat{\text{var}}(\hat{\beta}_{CO-(2)}) = \frac{\bar{P}'(1-\bar{P}')\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k-\bar{S}')^2 Y_{ijk}\bar{C}_{i(O)}}{[\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k-\bar{S}')^2 Y_{ijk}]^2}.$$

Consequently the trend test statistic for clustered ordinal outcomes is given by

$$\chi_{CO-(2)}^{2} = \frac{[\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k-\bar{S}')(\hat{P}_k-\bar{P}')Y_{ijk}]^2}{\bar{P}'(1-\bar{P}')\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k-\bar{S})^2 Y_{ijk}\bar{C}_{i(O)}}. \quad (3.11)$$

The statistic has all the properties of $\chi^2_{CO-(1)}$ in equation (3.9).

3.4.3 Extension of Weighted Least Squares Cochran-Armitage Test

Under the null hypothesis $H_0 : P_{ijk} = P_k = \bar{P}$, Y_{ijl} has a variance of $\sigma^2 = \bar{P}(1-\bar{P})$. Let V_{ij} represent the variance matrix for a single cluster given by

$$V_{ij} = \sigma^2\{(1-\rho_O)I + \rho_O J\}$$

where I denotes a $m_{ij} \times m_{ij}$ identity matrix and J the $m_{ij} \times m_{ij}$ matrix all of whose element are 1. Let V be a block-diagonal variance matrix with non-zero $m_{ij} \times m_{ij}$ blocks V_{ij}. We denote W as the $M \times M$ weight matrix for the WLS approach, where

$$W = V^{-1}.$$

From the linear model in (3.8), the WLS estimator of β_C is given by

$$\hat{\beta}_{CO-WLS} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k-\tilde{S}')(\hat{P}_k-\tilde{P}')\frac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}(S_k-\tilde{S}')^2\frac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}}.$$

where

$$\tilde{P}' = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}\hat{P}_K \dfrac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}\dfrac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}} \text{ and } \tilde{S}' = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}S_k \dfrac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}\dfrac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}}.$$

Here we use the superscript ' to distinguish $\tilde{P}'$ and $\tilde{S}'$ for ordinal outcomes with $\tilde{P}$ and $\tilde{S}$ for binary outcome. Under the null hypothesis, the corresponding variance estimator is given by

$$\widehat{\text{var}}(\hat{\beta}_{CO-WLS}) = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{k=1}^{K}\frac{\tilde{P}'(1-\tilde{P}')}{(S_k-\tilde{S}')^2\dfrac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}}.$$

Consequently, the WLS trend test statistic for clustered ordinal outcome data is given by

$$\chi_{CO-WLS}^{\ 2} = \frac{[\sum_{k=1}^{K}(S_k-\tilde{S}')(\hat{P}_k-\tilde{P}')\sum_{i=1}^{2}\sum_{j=1}^{n_i}\dfrac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}]^2}{\tilde{P}'(1-\tilde{P}')\sum_{k=1}^{K}(S_k-\tilde{S}')^2\sum_{i=1}^{2}\sum_{j=1}^{n_i}\dfrac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}}. \qquad (3.13)$$

In the special case of $\rho_O = 0$, χ^2_{CO-WLS} reduces to the regular C-A test. When $m_{ij} = m$, χ^2_{CO-WLS} reduces to

$$\chi_{CO-WLS}^{\ 2} = \frac{[\sum_{k=1}^{K}(S_k-\tilde{S}')(\hat{P}_k-\tilde{P}')\sum_{i=1}^{2}\sum_{j=1}^{n_i}\dfrac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}]^2}{\tilde{P}'(1-\tilde{P}')\sum_{k=1}^{K}(S_k-\tilde{S}')^2\sum_{i=1}^{2}\sum_{j=1}^{n_i}\dfrac{Y_{ijk}}{1+(m_{ij}-1)\hat{\rho}_O}} = \frac{\chi^2}{1+(m-1)\hat{\rho}_O}. \qquad (3.14)$$

When the outcomes have only two categories and the cluster size is constant, χ^2_{CO-WLS} reduces to Donner and Donald (1988)'s adjusted Pearson chi-square test statistic.

In section 3.3.3, we reviewed the close relationship between the WLS approach with other sophisticated methods. We also built the approximate equivalence between χ^2_{CB-WLS} and the score test statistic using the GEE approach in a binary logistic regression. We will further derive the relationship between these two statistics for clustered ordinal outcomes in Chapter 5.

3.5 Discussion

Although it is necessary to also develop methodology to address very general questions, including the analysis of covariates, it remains helpful to derive direct adjustments to regular methods which are simple and easily implemented. The adjustment approaches presented here can be calculated using any standard computer software designed for independent data. Furthermore, the principle underlying the adjustments in this chapter has been applied to a variety of problems involving clustered data. For example, Donner and Banting (1989) and Rao and Scott (1992) adopted it to the Pearson chi-square statistic and the Mantel-Haenszel chi-square statistic.

An assumption behind the adjustment approaches proposed in section 3.3 for clustered binary outcomes is that the correlation between any two observations in the same cluster is exchangeable, or the average of correlations among observations in a cluster remains constant across clusters. The random allocation of clusters in cluster randomization trials assures this assumption is reasonable, at least under H_o. Similar assumption may be guaranteed for adjustment approaches for clustered ordinal outcomes, at least under H_o.

In addition to $\overline{C}_i$, one may estimate the design effect by regarding the success rate $\hat{P}_i$ as a ratio rather than as a proportion (Cochran, 1977). Then the design effect estimator d_i is defined as the ratio of the estimated variance of $\hat{P}_i$ to its estimated variance assuming independent data. Rao and Scott (1992) proposed an adjusted Cochran-Armitage test by using d_i. However, their method estimates the design effect separately in each group so it is well-suited for non-randomization trials. Therefore, $\overline{C}_i$, rather than d_i, is adopted to account for the clustering effect in this thesis since our research interests focus on randomization trials.

Chapter 4

4 Marginal and cluster-specific models

4.1 Introduction

In Chapter 3, we proposed simple adjustments to Cochran-Armitage tests for comparisons between clustered ordinal outcomes from two groups. In this chapter, we present marginal and cluster-specific models, two typical modeling-based methods for the analysis of correlated categorical data. In particular, algebraic background is provided with emphases on the GEE and cluster-specific extension of proportional odds models with one single cluster-level covariate.

This chapter differs from the earlier work in two ways. Firstly, most earlier studies illustrated their methods using a general form of covariate structure and model links (e.g., Lipsitz et al., 1994; Rabe-Hesketh et al. 2002), or did not focus directly on cluster randomization trials (e.g., Hedeker, 2003; Raman and Hedeker, 2005). Very few studies gave explicit technical results in the context of a single cluster-level covariate and the cumulative logit link, the link most commonly used for ordinal outcomes arisen from epidemiologic studies. However, algebraic formulae related to modelling methods for correlated ordinal outcomes are more complicated than those for binary outcomes. As such, the illustration of modelling approaches for clustered ordinal outcomes requires more explicit details. In this chapter, we present an analytic investigation of marginal and cluster-specific extensions of ordinal logistic regression models applicable to cluster randomization trials.

Secondly, few earlier studies described both fitting procedures and hypothesis testing. Rather their focus has been on either fitting approaches in the two models only (e.g., Agresti and Natarajan, 2001), or hypothesis testing only (e.g., Boos, 1992). Moreover, most existing model-dependent statistical tests were illustrated particularly for correlated binary data (e.g., Rotnitaky and Jewell, 1990). Extensions of them to correlated ordinal outcomes were implied only.

The remainder of the chapter is organized as follows. In section 4.2 we discuss the GEE extension of the proportional odds models including its robust Wald test and score test. Section 4.3 briefly discusses the cluster-specific extension of proportional odds models. Section 4.4 discusses relationships among the magnitudes of fixed effects parameters and the variances of their estimates as obtained from marginal and cluster-specific models. Section 4.5 presents ICC estimation in the two models.

4.2 GEE extension of proportional odds logistic regression

We introduced Lipsitz et al.'s (1994) GEE approach for analyses of correlated ordinal outcomes in section 1.5.4.2. Here we adapt their approach to proportional odds models with a single binary cluster-level covariate.

4.2.1 Model formulation

Following Lipsitz et al. (1994), we denote Z_{ijlk} as the cumulative indicator of a *K*-level ordinal outcome where $Z_{ijlk} = 1$ if $Y_{ijl} \le k$ or $Z_{ijlk} = 0$ if $Y_{ijl} > k$. Here $k = 1,2,...K$ and $l = 1,2,...,m_{ij}$ where m_{ij} denotes cluster size for the *ij*th cluster. Letting $\gamma_{ijlk} = P(Y_{ijl} \le k)$ be the cumulative probability that has the form $E(Z_{ijlk}) = \gamma_{ijlk}$, then $Z_{ijl} = [Z_{ijl1}, Z_{ijl2},..., Z_{ijl(K-1)}]'$ and $\gamma_{ijl} = [\gamma_{ijl1}, \gamma_{ijl2}..., \gamma_{ijl(K-1)}]'$. As such, we transform the ordinal score Y_{ijl} to a new set of $K-1$ binary indicators Z_{ijlk} with cumulative probabilities γ_{ijl} corresponding to the cumulative logit link.

A marginal model based on cumulative logits has the form

$$\log it[\gamma_{ijl}] = X_{ijl}\beta^{*}. \quad (4.1)$$

Here X_{ijl} denotes a $(K-1)\times K$ design matrix for the *l*th observation in the *ij*th cluster and $\beta^{*} = [\alpha_1, \alpha_2,..., \alpha_{K-1}, \beta]'$ denotes a $K \times 1$ parameter vector. The intercept parameter, α_k corresponds to the *k*th cumulative logit and it is increasing as *k* increases. The intervention effect parameter, β denotes the log(odds ratio) of the cumulative probabilities comparing the experimental to the control group. Since the model in (4.1)

has the same effects β for each logit, the cumulative odds ratio is also constant for each logit. In this study, we are interested in the intervention effect parameter β only.

4.2.2 Estimation and inference

Let $X_{ij} = [X'_{ij1}, X'_{ij2}, ..., X'_{ijm_{ij}}]'$ denote the $m_{ij}(K-1) \times K$ design matrix, $Z_{ij} = [Z'_{ij1}, Z'_{ijl2}, ..., Z'_{ijm_{ij}}]'$ the $m_{ij}(K-1)$ cumulative response vector, and $\gamma_{ij} = [\gamma'_{ij1}, \gamma'_{ij2}, ..., \gamma'_{ijm_{ij}}]'$ the cumulative probabilities for the *ij*th cluster. Let B_{ij} denote a $[m_{ij}(K-1)] \times [m_{ij}(K-1)]$ diagonal matrix with the marginal variances of the elements of Z_{ij}, $\gamma_{ijlk}(1-\gamma_{ijlk})$, on the main diagonal and zeros elsewhere. We further assume a $[m_{ij}(K-1)] \times [m_{ij}(K-1)]$ working covariance matrix V_{ij} for the *ij*th cluster, given by

$$V_{ij} = B_{ij}^{1/2} R_{ij} B_{ij}^{1/2} \qquad (4.2)$$

where R_{ij} is a $[m_{ij}(K-1)] \times [m_{ij}(K-1)]$ working correlation matrix. Then the diagonal blocks of V_{ij} is the $(K-1) \times (K-1)$ multinomial covariance matrix for Z_{ijl},

$$V_{ijl} = Diag[\gamma_{ijl}] - \gamma_{ijl}\gamma'_{ijl}. \quad (4.3)$$

The remaining elements of V_{ij} contain the covariance between pairs Z_{ijlk} and Z_{ijhg} ($l, h = 1,2,..., m_{ij}; k, g = 1,2,..., K-1$). Additionally, the true covariance matrix of Z_{ij} is given by

$$\text{cov}(Z_{ij}) = B_{ij}^{1/2} R_{ij}^{0} B_{ij}^{1/2} \quad (4.4)$$

where R_{ij}^{0} is a $[m_{ij}(K-1)] \times [m_{ij}(K-1)]$ true correlation matrix of the *ij*th cluster. The working covariance V_{ij} in equation (4.2) is identical to the true covariance $\text{cov}(Z_{ij})$ in (4.4) only when the working correlation matrix R_{ij} is identical to R_{ij}^{0}.

Lipsitz et al. (1994) derived a generalized estimating equation in the form of

$$T = \sum_{i=1}^{2}\sum_{j=1}^{n_i} U_{ij} = \sum_{i=1}^{2}\sum_{j=1}^{n_i} D_{ij}' V_{ij}^{-1}[Z_{ij} - \hat{\gamma}_{ij}] = 0 \qquad (4.5)$$

where $D_{ij} = \partial\gamma_{ij} / \partial\beta^* = B_{ij} X_{ij}$. These estimating equations have the same form as the likelihood equations for logistic regression models, except that definitions of D_{ij} , V_{ij}, Z_{ij} and γ_{ij} have special meaning and structures for correlated ordinal outcomes as presented above.

Using equation (4.5), a Fisher scoring algorithm was suggested to obtain $\hat{\beta}$ in conjunction with the estimated correlation parameters in the working correlation at each iteration procedure. Therefore, given a starting value for β^*, the *m*th iteration procedure is given by

$$\hat{\beta}^{*(m+1)} = \hat{\beta}^{*(m)} + [\sum_{i=1}^{2}\sum_{j=1}^{n_i} \hat{D}'^{(m)}_{ij} \hat{V}^{-1(m)}_{ij} \hat{D}^{(m)}_{ij}]^{-1} \sum_{i=1}^{2}\sum_{j=1}^{n_i} \hat{D}'^{(m)}_{ij} \hat{V}^{-1(m)}_{ij} (Z_{ij} - \hat{\gamma}_{ij}) . \qquad (4.6)$$

Here $\hat{D}^{(m)}_{ij}$ and $\hat{V}^{(m)}_{ij}$ are estimated by substituting $\hat{\beta}^*$ and the correlation estimators in R_{ij} at the *m*th step. In particular, Lipsitz et al. (1994) extended Liang and Zeger (1986)'s method of moments approach to estimate the correlation parameters. We further discuss this in section 4.5.

The sandwich (robust) covariance matrix of model parameter vector $\hat{\beta}^*$ can be shown to be

$$V_R = V_M V_0 V_M . \qquad (4.7)$$

Here V_M denotes the model-based covariance of $\hat{\beta}^*$, given by

$$V_M = [\sum_{i=1}^{2}\sum_{j=1}^{n_i} D_{ij}' V_{ij}^{-1} D_{ij}]^{-1} \qquad (4.8)$$

and

$$V_0 = \sum_{i=1}^{2}\sum_{j=1}^{n_i} D_{ij}' V_{ij}^{-1} \operatorname{cov}(Z_{ij}) V_{ij}^{-1} D_{ij}' = \sum_{i=1}^{2}\sum_{j=1}^{n_i} U_{ij} U_{ij}' . \quad (4.9)$$

Specifically, we are interested in testing the null hypothesis $H_0 : \beta = 0$. The estimator $\hat{\beta}$ is the *K*th element of $\hat{\beta}^*$. Correspondingly, its model-based variance estimator $\widehat{\operatorname{var}}_M(\hat{\beta})$ and robust variance estimator $\widehat{\operatorname{var}}_M(\hat{\beta})$ are the (*K*,*K*)th element of $\hat{V}_M$ and $\hat{V}_R$ respectively. As such, the model-based Wald test statistic under $H_0 : \beta = 0$ has the form

$$W_M = \frac{\hat{\beta}^2}{\widehat{\operatorname{var}}_M(\hat{\beta})} \sim \chi_1^2 \quad (4.10)$$

and the robust Wald test statistic is given by

$$W_R = \frac{\hat{\beta}^2}{\widehat{\operatorname{var}}_R(\hat{\beta})} \sim \chi_1^2 . \quad (4.11)$$

In chapter 6, we will evaluate these model-based test statistics by simulation using SAS procedures. Therefore we limit attention to those statistics routinely available in SAS.

Since only the independence working correlation is available for multinomial models in PROC GENMOD, the Wald test statistics in (4.10) and (4.11) are considered assuming independence working correlation only in this study.

Next we introduce the score test obtained from equation (4.5). Corresponding to H_0, we decompose the parameter vector β^* to $(\beta, \beta_{(1)}')'$, where β is the parameter being tested in the null hypothesis and $\beta_{(1)}$ is a $(K-1)\times 1$ parameter vector with elements α_k. Similarly, we decompose the generalized estimating equation (4.5) to $T = (T_{(0)}', T_{(1)}')'$, decompose V_M to four submatrices $A_{(00)}$ (i.e., $\operatorname{var}_M(\hat{\beta})$), $A_{(01)}$, $A_{(10)}$ and $A_{(11)}$, and decompose V_R to four submatrices $J_{(00)}$ (i.e., $\operatorname{var}_R(\hat{\beta})$), $J_{(01)}$, $J_{(10)}$ and $J_{(11)}$,

corresponding to β and $\beta_{(1)}$ .We first obtain the estimate $\tilde{\beta}_{(1)}$ under H_0 by solving $T_{(1)}$, then substitute $\tilde{\beta}_{(1)}$ into $T_{(0)}$ under H_0 to yield the numerator of the score statistic, i.e., $\tilde{T}_{(0)}$. Hence the model-based score test statistic for $H_0 : \beta = 0$ is given by

$$S_M = \tilde{T}_{(0)}^2 \widetilde{\text{var}}_M(\hat{\beta}) \sim \chi_1^2 \quad (4.17)$$

and the robust score test statistic is given by

$$S_R = \frac{\tilde{T}_{(0)}^2 \widetilde{\text{var}}_M^2(\hat{\beta})}{\widetilde{\text{var}}_R(\hat{\beta})} \sim \chi_1^2 . \quad (4.18)$$

Here $\widetilde{\text{var}}_M(\hat{\beta})$ and $\widetilde{\text{var}}_R(\hat{\beta})$ are obtained by substituting $\alpha = \tilde{\alpha}$ and $\beta = 0$ into $\widehat{\text{var}}_M(\hat{\beta})$ and $\widehat{\text{var}}_M(\hat{\beta})$ respectively.

Proofs giving the distributions of the above statistics are completely analogous to those for binary data. Thus we refer below to other authors who have provided corresponding results in the case of binary outcomes. In particular, Liang and Zeger (1986) showed that the robust Wald statistic asymptotically follows a chi-square distribution; Rotnitzky and Jewell (1990) demonstrated that the model-based Wald statistic has an asymptotic chi-square distribution if working correlations are correctly specified; furthermore, Rotnitzky and Jewell (1990) and Geys et al. (1999) provided the proof that the robust and the model-based score statistic have asymptotic chi-square distributions under H_0.

Rotnitzky and Jewell (1990) reported that robust Wald and score statistics for binary data may suffer unstable computational results if the cluster sizes are large and the number of clusters is small. This is because that the residual estimator of $\text{cov}(Z_{ij})$, the middle piece of sandwich estimators, is a quite variable estimator. Therefore simpler statistics, i.e., the model-based Wald or score statistic, may be used as an alternative if the working correlation matrix is correctly specified. However, their distributions under H_0 are complicated. Therefore adjustments have been proposed to model-based statistics so that they could be easily evaluated as approximate chi-square distributions (e.g., Rotnitzky

and Jewell, 1990; Geys et al, 1999). Their extensions to clustered ordinal data are outside the present scope of this research.

As discussed for GEE Wald tests, only the independence working correlation is available for multinomial models in PROC GENMOD. As such, we only consider the robust score tests assuming the independence working correlation in this study.

4.3 Cluster-specific extension of proportional odds logistic regression

4.3.1 Model formulation

In cluster-specific models, cluster effects are considered by adding a random effect term, which is commonly assumed to follow a normal distribution. Let the random variable $u_{ij} \sim N(0, \sigma^2)$ denote the random effect of the *ij*th cluster. Then a cluster-specific model for clustered ordinal outcomes with cumulative logit link is given by

$$\log it[P(Y_{ijl} \le k)] = X_{ijl}\beta^* + u_{ij}. \qquad (4.19)$$

Model (4.19) may be fit by maximum likelihood, which is discussed as follows. Here X_{ijl} denotes a $(K-1)\times K$ design matrix for the *l*th observation in the *ij*th cluster and $\beta^* = [\alpha_1, \alpha_2, ..., \alpha_{K-1}, \beta]'$ denotes a $K \times 1$ parameter vector. The intercept parameter, α_k corresponds to the *k*th cumulative logit and it is increasing as *k* increases. The intervention effect parameter, β denotes the log(odds ratio) of the cumulative probabilities comparing the experimental to the control group.

Let Y_{ij} denote a $m_{ij} \times 1$ response vector of scores for the *ij*th cluster over the set of m_{ij} observations. The likelihood function of Y_{ij} conditional on the random effects has the form

$$l(Y_{ij} \mid u_{ij}, \beta^*) = \prod_{l=1}^{m_{ij}} \prod_{k=1}^{K-1} p_{ijlk}^{Y_{ijlk}}. \qquad (4.20)$$

Here $Y_{ijlk} = 1$ if Y_{ijl} falls into the kth category and 0 otherwise, and $p_{ijlk} = P(Y_{ijlk} = 1 \mid x_{ijl}, u_{ij})$. We further define p_{ijlk} as a difference of cumulative probabilities with the inverse cumulative logits link. That is,

$$p_{ijlk} = \gamma_{ijlk} - \gamma_{ijl(k-1)} = \frac{1}{1+\exp(X_{ijl(k-1)}\beta^* + u_{ij})} - \frac{1}{1+\exp(X_{ijlk}\beta^* + u_{ij})}.$$

The likelihood function of the *ij*th cluster after integrating out the random effects is given by

$$h(Y_{ij}) = \int_{-\infty}^{+\infty} l(Y_{ij} \mid u_{ij}, \beta^*)\phi(u_{ij}, \sigma^2)du_{ij} \quad (4.21)$$

where $\phi(0, \sigma^2)$ represents the normal density function of u_{ij}.Then a full likelihood function is given by

$$L = \prod_{i=1}^{2}\prod_{j=1}^{n_i} h(Y_{ij}). \quad (4.22)$$

However, the integrals in (4.22) don't have a closed form expression and numerical approximations are required, which are discussed in the next section.

4.3.2 Estimation and inference

Agresti and Natarajan (2001) reviewed approximation methods for the likelihood function in (4.22), and they suggested that the best method is Gauss-Hermite quadrature.

As an alternative, the PQL approach (Breslow and Clayton, 1993) has also been commonly used in cluster-specific models. Bellamy et al. (2005) argued that it may be a reasonable choice for cluster randomization trials where there are small numbers of large clusters. However, this approach tends underestimate regression coefficients as well as variance components for binary outcome data (Breslow and Clayton, 1993; Jang and Lim, 2006). Liu and Agresti (2005) claimed that similar problems may exist for ordinal outcome data.

As such, only the Gauss-Hermite quadrature approximation approach is considered for estimation of cluster-specific models. However, the Gauss-Hermite quadrature approach is not dealt with in detail since our interest focuses on marginal models, especially the GEE approach (see section 4.1).

In Gauss-Hermite quadrature, the likelihood function is approximated by a weighted sum of a specified number of quadrature points Q_h with the weight w_h. Optional choices of points and weights have been reported. For example, Stroud and Sechrest (1966) list optimal points and weights for the standard normal univariate distribution.

The Gauss-Hermite quadrature approximation of (4.21) is the weighted sum

$$h(Y_{ij}) \approx \sum_{h=1}^{q} l(Y_{ij} \mid Q_h, \beta^*) w_h \quad (4.24)$$

Here q is the number of quadrature points. The accuracy of the approximation increases as q becomes larger. The ML estimates of β^* and σ and their variance estimates can then be obtained by evaluating the approximated likelihood function using standard algorithms.

Then the corresponding Wald test statistic for $H_0 : \beta_{CS} = 0$ has the form

$$W_{CS} = \frac{\hat{\beta}_{CS}^{\ 2}}{\widehat{\text{var}}_{CS}(\hat{\beta}_{CS})} \sim \chi_1^2, \quad (4.25)$$

where β_{CS} denotes the intervention effect parameter in the cluster-specific model (4.19). Since only the Wald test is available for cluster-specific in SAS procedures, we do not consider the score tests in cluster-specific model.

In addition to the Gauss-Hermite quadrature, the adaptive version of Gauss-Hermite quadrature has been proposed. It increases the efficiency of the ordinary Gauss-Hermite quadrature so that fewer quadrature points are required. For details one could refer to Liu and Pierce (1994), Pinheiro and Chao (2006), and Rabe-Hesketh et al. (2005).

4.4 Relationship between Marginal and Cluster-specific Models

Although both marginal and cluster-specific models can be viewed as extensions of generalized linear models to correlated data, they have different interpretations, estimating methods and in general, yield different results. In this section, we discuss relationships between the two models in terms of magnitudes and standard errors of the estimated intervention effect parameter.

Let β_M and β_{CS} denote intervention effect parameters in marginal and cluster-specific models respectively. When the outcomes are binary, Zeger et al. (1988) showed that the approximate relationship between the two parameters has the form

$$\beta_M \approx \left[\left(16\sqrt{3}/15\pi\right)^2 \sigma^2 + 1\right]^{-1/2} \beta_{CS} \quad (4.27)$$

under the assumption that the random effect distribution is normal. Here σ^2 represents the variance of the random effect, i.e., $u_{ij} \sim N(0, \sigma^2)$. In addition, Neuhaus et al. (1991) derived a similar relationship which is valid under any random effect distribution. They used a first-order Taylor series approximation about $\beta_{CS} = 0$ and obtain

$$\beta_M \approx [1 - \rho(0)]\beta_{CS} . \quad (4.28)$$

Please note that equation (4.28) was derived under the assumption of any random effect distribution, and $\rho(0)$ is the intracluster correlation obtained under the null hypothesis $\beta_{CS} = 0$.

Since σ^2 is a function of the intracluster correlation ρ, (4.27) and (4.28) illustrate a qualitatively similar relationship between β_M and β_{CS}. Both of them show that for clustered binary data β_M is smaller than β_{CS} and the discrepancy between β_M and β_{CS} increases as the intracluster correlation increases. However, in community intervention trials, the discrepancy between β_M and β_{CS} would be very little since ICCs in community intervention trials tend to be near zero (see section 6.2).

Neuhaus (1993) also discussed the relationship between variances of $\hat{\beta}_M$ and $\hat{\beta}_{CS}$ for clustered binary outcomes. Under the null hypothesis, the relationship between these variances assuming an independence working correlation structure is given by

$$\text{var}(\hat{\beta}_M) \approx \frac{1-\rho(0)}{1+\rho(0)} \text{var}(\hat{\beta}_{CS}), \quad (4.29)$$

and with the exchangeable working correlation structure, is given by

$$\text{var}(\hat{\beta}_M) \approx (1-\rho(0))^2 \text{var}(\hat{\beta}_{CS}). \quad (4.30)$$

In community intervention trials, small values of ICCs would decrease the difference between variances of estimated regression coefficients from the two models.

For clustered ordinal outcomes, Ten Have et al. (1996) extended Zeger et al.'s (1988) approach and showed that the relationship between the magnitudes of fixed effect estimates for clustered ordinal outcomes parallels that reported for clustered binary outcomes. As such, in community intervention trials β_M would be slightly smaller than β_{CS} as the ICCs tend to be near zero.

However, the analytical derivation of the relationship between the variances is more complicated. Therefore Ten Have et al. (1996) compared variances using real data. They concluded that the relationship between the variances arising from the two models for binary outcomes does not hold for ordinal outcomes.

Ten Have et al.'s (1996) conclusions are based on empirical comparisons. Although example datasets are useful for illustration purposes, simulation studies are needed to provide more evidence under varying parameter combinations (e.g., varying number of clusters and ICCs) to assess the performance of different statistical techniques. In chapter 6, we will examine Ten Have et al.'s (1996) conclusions using simulation.

4.5 ICC estimation

In section 2.3, we reviewed ICC estimating methods for clustered ordinal outcomes. In this section, we further discuss estimation of the ICC under GEE and cluster-specific extensions of proportional odds logistic regressions.

4.5.1 ICC estimation under marginal models

As introduced in section (2.3.2), Lipsitz et al. (1994) derived the moment ICC estimator in GEE approach for correlated ordinal outcomes. For an exchangeable correlation structure, the ICC estimator in model (4.1) is given by

$$\hat{\rho}_{GEE} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{t>s}\hat{e}_{ijs}\hat{e}'_{ijt}}{[\sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{1}{2}m_{ij}(m_{ij}-1)]-3}. \quad (4.31)$$

Here the $m_{ij}(K-1)$ residual vector $\hat{e}_{ijl} = \hat{B}_{ijl}^{-\frac{1}{2}}[Z_{ijl} - \hat{\gamma}_{ijl}]$ corresponds to the cumulative logit links and B_{ijl} is the submatrix of matrix B_{ij} corresponding to the *l*th individual in the *ij*th cluster. When there is substantial variation in cluster size, the precision of the ICC estimator in (4.31) may not be optimal since it gives too much weight to large groups (Donner, 1986).

4.5.2 ICC estimation in cluster-specific models

Rodriguez and Goldman (2001) extended the classical derivation of the ANOVA ICC for continuous data to binary outcomes based on a latent-variable formulation of generalized linear mixed models. Agreti (2010, page 283-284) further discussed the ICC for correlated ordinal outcomes using the latent variable formulation in cluster-specific models.

Given the random effect term $u_{ij} \sim N(0,\sigma^2)$ and the error term $\varepsilon_{ijl} \sim N(0,\sigma_{ijl}^2)$, the ICC estimator in model (4.19) assuming a common correlation structure has the form

$$\rho_{CS} = \frac{\sigma^2}{\sigma^2 + \sigma_{ijl}^2} = \frac{\sigma^2}{\sigma^2 + \pi^2/3}. \quad (4.32)$$

This implies a nonnegative ICC among clustered observations and it tends to increase as the variance σ^2 of the random effect increases.

4.6 Summary

In this chapter we presented marginal and cluster-specific models which will be investigated by simulation. In section 4.2, we introduced the GEE extension of proportional odds logistic regressions. In section 4.3, we briefly introduced cluster-specific extensions of proportional odds logistic regressions. In section 4.4, we discussed relationships between marginal and cluster-specific models. In section 4.5, the estimation of the ICC under the two models was introduced.

Chapter 5

5 Adjustments to the small-sample performance of GEE

5.1 Introduction

As reviewed in section 1.5.4, correction and modification strategies have been proposed to improve the small-sample performance of the GEE approach for correlated binary data. However, their extensions to correlated ordinal data have not been considered. In this chapter, we develop modified GEE procedures for ordinal outcome data to improve small-sample performance of hypothesis tests.

The specific objective of this chapter is to algebraically extend correction and modification strategies developed for binary outcome data to ordinal outcome data. For convenience, we simply classify adjustments to the robust Wald test into two categories: one based on bias-corrections of the sandwich estimator and the other based on degrees-of-freedom adjustments for the test distributions. As such, we consider five bias-corrected approaches and four degree-of-freedom approaches to the robust Wald test and one modified score test. We list all test statistics in Table 5.1. The subscript 'BC' denotes bias-corrected and 'df' denotes degree-of-freedom adjusted. In addition, the subscript 'M' denotes model-based and 'R' denotes robust.

Most attention given to small samples adjustments in marginal models applicable to cluster randomization trials has focused on a single cluster-level binary covariate and cumulative logit links. Although the test statistics presented in this chapter are derived similarly to those for binary outcomes, detailed attention is given to technical issues that arise in the case of ordinal data.

The remainder of the chapter is organized as follows. In section 5.2, we adapt small-sample adjustments to the robust Wald tests for ordinal outcomes. In section 5.3, we present modified score tests for ordinal outcomes.

Table 5.1: Small-sample adjustments to Wald and Score tests in Chapter 5

Test statistic	Test name	Formula	Test distribution	Equation
W_M	Model-based Wald test	$\hat{\beta}^2 / \widehat{\text{var}}_M(\hat{\beta})$	χ^2_1	Equation (4.10)
W_R	Robust Wald test	$\hat{\beta}^2 / \widehat{\text{var}}_R(\hat{\beta})$	χ^2_1	Equation (4.11)
W_{BC1}	Bias-corrected Wald test: Approach 1	$\hat{\beta}^2 / \widehat{\text{var}}_{BC1}(\hat{\beta})$	χ^2_1	Equation (5.7)
W_{BC2}	Bias-corrected Wald test: Approach 2	$\hat{\beta}^2 / \widehat{\text{var}}_{BC2}(\hat{\beta})$	χ^2_1	Equation (5.10)
W_{BC3}	Bias-corrected Wald test: Approach 3	$\hat{\beta}^2 / \widehat{\text{var}}_{BC3}(\hat{\beta})$	χ^2_1	Equation (5.16)
W_{BC4}	Bias-corrected Wald test: Approach 4	$\hat{\beta}^2 / \widehat{\text{var}}_{BC4}(\hat{\beta})$	χ^2_1	Equation (5.19)
W_{BC5}	Bias-corrected Wald test: Approach 5	$\hat{\beta}^2 / \widehat{\text{var}}_{BC5}(\hat{\beta})$	χ^2_1	Equation (5.23)
W_{df1}	Degrees-of-freedom-adjusted Wald test: Approach 1	$\hat{\beta}^2(N-K) / \widehat{\text{var}}_R(\hat{\beta})N$	χ^2_1	Equation (5.24)
W_{df2}	Degrees-of-freedom-adjusted Wald test: Approach 2	$\hat{\beta}^2 / \widehat{\text{var}}_R(\hat{\beta})$	$F_{1,N-K}$	Equation (5.25)
W_{df3}	Degrees-of-freedom-adjusted Wald test: Approach 3	$\hat{\beta}^2 / \widehat{\text{var}}_R(\hat{\beta})$	$F_{1,d}$	Equation (5.26)
W_{df4}	Degrees-of-freedom-adjusted Wald test: Approach 4	$\hat{\beta}^2 / \widehat{\text{var}}_R(\hat{\beta})$	$F_{1,d'}$	Equation (5.36)
S_M	Model-based score test	$\tilde{T}^2_{(0)} \widetilde{\text{var}}_M(\hat{\beta})$	χ^2_1	Equation (4.17)
S_R	Robust score test	$\tilde{T}^2_{(0)} \widetilde{\text{var}}^2_M(\hat{\beta}) / \widetilde{\text{var}}_R(\hat{\beta})$	χ^2_1	Equation (4.18)
S_{BC}	Modified robust score test	$\tilde{T}^2_{(0)} \widetilde{\text{var}}^2_M(\hat{\beta})N / \widetilde{\text{var}}_R(\hat{\beta})(N-1)$	χ^2_1	Equation (5.38)

5.2 Adjustments to the Wald test

5.2.1 Bias-corrected approaches

Approach 1.

To calculate the sandwich estimator in equation (4.7), the estimated residuals $\hat{r}_{ij} = Z_{ij} - \hat{\gamma}_{ij}$ are commonly used to estimate $\text{cov}(Z_{ij})$, i.e.,

$$\widehat{\text{cov}}(Z_{ij}) = \hat{r}_{ij}\hat{r}'_{ij} = (Z_{ij} - \hat{\gamma}_{ij})(Z_{ij} - \hat{\gamma}_{ij})'. \quad (5.1)$$

However, the residuals $\hat{r}_{ij}$ tend to be too small so $\hat{r}_{ij}\hat{r}'_{ij}$ is a biased estimator of $\text{cov}(Z_{ij})$. To derive the approximate bias of the residual estimator, Mancl and DeRouen (2001) considered a first-order Taylor expansion of the residual $\hat{r}_{ij}$, given by

$$\hat{r}_{ij} = r_{ij} + \frac{\partial r_{ij}}{\partial \beta'^{*}}(\hat{\beta}^{*} - \beta^{*}), \quad (5.2)$$

and the first-order approximation

$$\hat{\beta}^{*} - \beta^{*} \approx V_M \sum_{i=1}^{2}\sum_{j=1}^{n_i} D_i' V_i^{-1}(Z_{ij} - \gamma_{ij}). \quad (5.3).$$

Substituting (5.3) into (5.2), we derive the expectation of $\hat{r}_{ij}\hat{r}'_{ij}$ as

$$E(\hat{r}_{ij}\hat{r}'_{ij}) \approx (I_{ij} - H_{ij})\,\text{cov}(Z_{ij})(I_{ij} - H_{ij})' + \sum_{i=1}^{2}\sum_{d\neq j} H_{id}\,\text{cov}(Z_{ij})H'_{id}, \quad (5.4)$$

where $H_{ij} = D_{ij}V_m D'_{ij}V_{ij}^{-1}$ is an expression for the leverage of the *ij*th cluster (Preisser and Qaqish, 1996), I_{ij} is the identity matrix with the same dimension as H_{ij}, and the summation $\sum_{d\neq j}$ in (5.4) is over all $d = 1,2,...,n_i \neq j$. By definition, the elements of H_{ij} are between zero and one, so we assume that the contribution to the bias of the sum

in Equation (5.4) is negligible. As such, the expectation of $\hat{r}_{ij}\hat{r}'_{ij}$ could be approximated by

$$E(\hat{r}_{ij}\hat{r}'_{ij}) \approx (I_{ij} - H_{ij})\,\mathrm{cov}(Z_{ij})(I_{ij} - H_{ij})'. \quad (5.5)$$

A bias-corrected sandwich variance estimator then has the form

$$V_{BC1} = V_M \left(\sum_{i=1}^{2} \sum_{j=1}^{n_i} D'_{ij} V_{ij}^{-1} (I_{ij} - H_{ij})^{-1} r_{ij} r'_{ij} (I_{ij} - H'_{ij})^{-1} V_{ij}^{-1} D_{ij} \right) V_M, \quad (5.6)$$

where $E(V_{BC1}) = \mathrm{var}(\beta^*)$.

Letting $\mathrm{var}_{BC1}(\hat{\beta})$ be the (K,K)th element in V_{BC1}, denoting the corrected sandwich variance of $\hat{\beta}$, the corresponding bias-corrected Wald test under the null hypothesis $H_0 : \beta = 0$ has the form

$$W_{BC1} = \frac{\hat{\beta}^2}{\widehat{\mathrm{var}}_{BC1}(\hat{\beta})} \sim \chi_1^2 \quad . \quad (5.7)$$

Approach 2.

Kauermann and Carroll (2001) proposed an alternative bias-corrected sandwich estimator for binary outcomes. One of its distinctions as compared to Mancl and DeRouen's (2001) approach is that it assumes a correctly specified working covariance, i.e., $V_{ij} = \mathrm{cov}(Z_{ij})$ and $E(r_{ij}r'_{ij}) = V_{ij}$. Also, it does not drop the summation term in (5.4). Next we derive an alternative bias correction to ordinal data based on Kauermann and Carroll's (2001) work. However, this new approach simplifies Kauermann and Carrol's method and leads to a result more comparable with other corrections.

We assume a correctly specified working correlation matrix, and substitute the first-order Taylor expansions of $\hat{\beta}^* - \beta$ into equation (5.2). Thus the expectation of $\hat{r}_{ij}\hat{r}'_{ij}$ is approximated by

$$E(\hat{r}_{ij}\hat{r}'_{ij}) = \text{cov}(Z_{ij})\{I_{ij} - H_{ij}\}. \quad (5.8)$$

The corrected residual estimator may then be written as $(I_{ij} - H_{ij})^{-1/2}\hat{r}_{ij}$. A bias-corrected sandwich estimator for clustered ordinal outcomes is then given by

$$V_{BC2} = V_M\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i} D'_{ij}V_{ij}^{-1}(I_{ij} - H_{ij})^{-1/2}r_{ij}r'_{ij}(I_{ij} - H_{ij})^{-1/2}V_{ij}^{-1}D_{ij}\right)V_M, \quad (5.9)$$

where $E(V_{BC2}) = \text{var}(\beta^*)$. Letting $\widehat{\text{var}}_{BC2}(\hat{\beta})$ be the (*K*,*K*)th element in $\hat{V}_{BC2}$, the corresponding bias-corrected Wald test under the null hypothesis $H_0 : \beta = 0$ has the form

$$W_{BC2} = \frac{\hat{\beta}^2}{\widehat{\text{var}}_{BC2}(\hat{\beta})} \sim \chi_1^2. \quad (5.10)$$

For clustered binary outcomes, Lu et al. (2007) reported that the Mancl and DeRouen estimator overestimates the true variance while the Kauermann and Carrol estimator reduces this overcorrection. Also, when cluster sizes are small (e.g. $m_{ij} < 10$), the Kauermann and Carrol estimator is preferred in terms of confidence interval coverage. However, when cluster sizes are moderate to large, the Mancl and DeRouen estimator performs better than the Kauermann and Carroll estimator in terms of coverage, even in trials with as few as 10 clusters (Lu et al., 2007). Note that these conclusions are reported for clustered binary outcomes only. In chapter 6, we will investigate the performance of $\widehat{\text{var}}_{BC1}(\hat{\beta})$ and $\widehat{\text{var}}_{BC2}(\hat{\beta})$ to find if similar results hold for clustered ordinal outcomes.

Approach 3.

We derived $\widehat{\text{var}}_{BC1}(\hat{\beta})$ and $\widehat{\text{var}}_{BC2}(\hat{\beta})$ by combining a first order Taylor expansion of residuals r_{ij} together with the Taylor expansion of β^*. Similarly, we could combine a Taylor expansion of estimating equations U_{ij} together with expansion of β^* to derive the bias of the sandwich estimator (Fay and Graubard, 2001). Substituting

$$\hat{\beta}^* - \beta^* \approx V_M \sum_{i=1}^{2} \sum_{j=1}^{n_i} D'_i V_i^{-1} (Z_{ij} - \gamma_{ij})$$

into the Taylor expansion of estimating equations

$$U_{ij} \approx \hat{U}_{ij} - \frac{\partial U_{ij}}{\partial \beta'} (\hat{\beta}^* - \beta^*), \quad (5.11)$$

we obtain

$$\begin{aligned} & E(\hat{U}_{ij} \hat{U}'_{ij}) \\ & = \operatorname{cov}(U_{ij}) - \operatorname{cov}(U_{ij}) V_M D'_{ij} V_{ij}^{-1} D_{ij} - D'_{ij} V_{ij}^{-1} D_{ij} V_M \operatorname{cov}(U_{ij}) \\ & + D'_{ij} V_{ij}^{-1} D_{ij} V_M \{ \sum_{i=1}^{2} \sum_{j=1}^{m_{ij}} \operatorname{cov}(U_{ij}) \} V_M D'_{ij} V_{ij}^{-1} D_{ij} \end{aligned} \quad (5.12)$$

Assuming a correctly specified working correlation matrix, we have

$$\operatorname{cov}(U_{ij}) = E(U_{ij} U'_{ij}) = D'_{ij} V_{ij}^{-1} D_{ij}. \quad (5.13)$$

Replacing $\sum_{i=1}^{2} \sum_{j=1}^{n_i} \operatorname{cov}(U_{ij})$ in (5.12) by (5.13) yields

$$E(\hat{U}_{ij} \hat{U}'_{ij}) \approx \operatorname{cov}(U_{ij})(I_{ij} - \Psi_{ij}) \quad (5.14)$$

where $\Psi_{ij} = D'_{ij} V_{ij}^{-1} D_{ij} V_m$. Since it is possible that $I_{ij} - \Psi_{ij}$ is not a symmetric matrix, we may be unable to use $(I_{ij} - \Psi_{ij})^{-1/2}$ to correct the bias of the approximation in (5.16). As Fay and Graubard (2001) proposed, we therefore select a constant b and define Δ_{ij} as a $K \times K$ diagonal matrix with ddth element equal to $\{1 - \min(b, [\psi_{ij}]_{dd})\}^{-1/2}$, where $b < 1$. Here we refer the choice of the constant b to Fay and Graubard (2001). A bias-corrected variance estimator is then given by

$$V_{BC3} = V_M \left[\sum_{i=1}^{2} \sum_{j=1}^{n_i} \operatorname{cov}(U_{ij} U'_{ij}) \right] V_M = V_M \left(\sum_{i=1}^{2} \sum_{j=1}^{n_i} \Delta_{ij} D'_{ij} V_{ij}^{-1} r_{ij} r'_{ij} V_{ij}^{-1} D_{ij} \Delta_{ij} \right) V_M, \quad (5.15)$$

where $E(V_{BC3}) = \text{var}(\beta^*)$ assuming a correctly specified working correlation.

Letting $\widehat{\text{var}}_{BC3}(\hat{\beta})$ be the (*K*,*K*)th element in $\hat{V}_{BC3}$, the corresponding bias-corrected Wald test under the hypothesis $H_0 : \beta = 0$ has the form

$$W_{BC3} = \frac{\hat{\beta}^2}{\widehat{\text{var}}_{BC3}(\hat{\beta})} \sim \chi_1^2 . \quad (5.16)$$

Approach 4.

Alternatively, we combine derivation procedures for $\widehat{\text{var}}_{BC1}(\hat{\beta})$ and $\widehat{\text{var}}_{BC3}(\hat{\beta})$ and develop another corrected sandwich estimator. In particular, we decompose the summation term of (5.12) and re-express the approximation (5.12) as

$$\begin{aligned} & E(\hat{U}_{ij}\hat{U}'_{ij}) \\ & = \text{cov}(U_{ij}) - \text{cov}(U_{ij})V_M D'_{ij} V_{ij}^{-1} D_{ij} - D'_{ij} V_{ij}^{-1} D_{ij} V_M \text{ cov}(U_{ij}) + D'_{ij} V_{ij}^{-1} D_{ij} V_M \text{ cov}(U_{ij}) V_M D'_{ij} V_{ij}^{-1} D_{ij} \\ & + D'_{ij} V_{ij}^{-1} D_{ij} V_M \sum_{i=1}^{2} \sum_{d \neq j} \text{cov}(U_{id}) V_M D'_{ij} V_{ij}^{-1} D_{ij} \end{aligned}$$

(5.17)

Neglecting the summation term in (5.17), the expectation of $\hat{U}_{ij}\hat{U}'_{ij}$ has the form

$$E(\hat{U}_{ij}\hat{U}'_{ij}) \approx (I_{ij} - \Psi_{ij})\text{cov}(U_{ij})(I_{ij} - \Psi_{ij})' .$$

Then a bias-corrected sandwich estimator is given by

$$V_{BC4} = V_M \left(\sum_{i=1}^{2} \sum_{j=1}^{n_i} (I_{ij} - \Psi_{ij})^{-1} D'_{ij} V_{ij}^{-1} r_{ij} r'_{ij} V_{ij}^{-1} D_{ij} (I_{ij} - \Psi_{ij})^{-1} \right) V_M \quad (5.18)$$

where $E(V_{BC3}) = \text{var}(\beta^*)$. Letting $\widehat{\text{var}}_{BC4}(\hat{\beta})$ be the (*K*,*K*)th element in $\hat{V}_{BC4}$, the corresponding bias-corrected Wald test under the hypothesis $H_0 : \beta = 0$ has the form

$$W_{BC4} = \frac{\hat{\beta}^2}{\widehat{\text{var}}_{BC4}(\hat{\beta})} \sim \chi_1^2. \quad (5.19)$$

In summary, derivations of $\widehat{\text{var}}_{BC2}(\hat{\beta})$ and $\widehat{\text{var}}_{BC3}(\hat{\beta})$ require an assumption of a correctly specified correlation matrix, and do not drop the summation term in (5.4) and (5.17) respectively. In contrast, derivations of $\widehat{\text{var}}_{BC1}(\hat{\beta})$ and $\widehat{\text{var}}_{BC4}(\hat{\beta})$ do not require this assumption, but they drop the summation term in (5.4) and (5.17) respectively.

In addition, $\widehat{\text{var}}_{BC4}(\hat{\beta})$ gives a larger variance estimator than $\widehat{\text{var}}_{BC3}(\hat{\beta})$ since the diagonal elements in Δ_{ij} are between 0 to 1. As such, based on conclusions derived from the binary case (Lu et al., 2007), V_{BC4} may overestimate the true variance. However, when the cluster size is moderate to large, the impact of overcorrection of $\widehat{\text{var}}_{BC4}(\hat{\beta})$ may counteract the high variability of the sandwich estimator. This is also one of motivations for deriving W_{BC4}. In the simulation study, we will further compare these adjustments to confirm our algebraic results.

Approach 5.

We also extend the Pan (2001) modification of the sandwich estimator to ordinal data. Pan (2001) reported that the residual estimator $r_{ij}r'_{ij}$ is not an optimal estimator of $\text{cov}(Z_{ij})$ in terms of consistency and efficiency since it is based on observations from only one cluster. Instead of $r_{ij}r'_{ij}$, he used Liang and Zeger (1986)'s estimator by pooling information across all clusters.

Following Liang and Zeger (1986), we could estimate the unspecified correlation R_0 in equation (4.4) by

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i} B_{ij}^{-1/2}(Z_{ij}-\gamma_{ij})(Z_{ij}-\gamma_{ij})' B_{ij}^{-1/2} / N \ . \quad (5.20)$$

Then from equation (4.4) we have

$$\mathrm{cov}(Z_{ij}) = B_{ij}^{1/2}\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i} B_{ij}^{-1/2}(Z_{ij}-\gamma_{ij})(Z_{ij}-\gamma_{ij})' B_{ij}^{-1/2} / N\right)B_{ij}^{1/2} \ . \quad (5.21)$$

Substituting (5.21) into V_R, the bias-corrected sandwich variance estimator is given by

$$V_{BC5} = V_M\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i} D'_{ij}V_{ij}^{-1}B_{ij}^{1/2}\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i} B_{ij}^{-1/2}(Z_{ij}-\gamma_{ij})(Z_{ij}-\gamma_{ij})' B_{ij}^{-1/2} / N\right)B_{ij}^{1/2}V_{ij}^{-1}D_{ij}\right)V_M \ . \quad (5.22)$$

Letting $\widehat{\mathrm{var}}_{BC5}(\hat{\beta})$ be the (K,K)th element in $\hat{V}_{BC5}$, the corresponding bias-corrected Wald test under the hypothesis $H_0 : \beta = 0$ has the form

$$W_{BC5} = \frac{\hat{\beta}^2}{\widehat{\mathrm{var}}_{BC5}(\hat{\beta})} \sim \chi_1^2 \ . \quad (5.23)$$

Proofs giving the chi-square distribution of the Wald statistic (5.23) are not as straightforward as those for the other four corrected statistics. We refer detailed proof procedures to Pan (2001).

5.2.2 Degrees-of-freedom adjusted approaches

In the previous section, we derived five bias-corrected sandwich variance estimators for robust Wald tests. In this section, we present four adjustments to the Wald tests in terms of degrees-of-freedom. We start with the simplest one.

Approach 1.

Hinkley (1977) and MacKinnon and White (1985) proposed a modification for heteroskedasticity-consistent variance estimators for continuous outcomes. Mancl and DeRouen (2001) adapted this approach for binary outcomes, what they called the degree-of-freedom approach, by multipling the sandwich estimator by a factor $N/(N-P)$.

We adapt this approach for ordinal data by using the same factor. Here P is equal to K, which denotes the number of categories of responses. Consequently, the corresponding Wald test statistic for the hypothesis $H_0 : \beta = 0$ is given by

$$W_{df1} = \frac{N-K}{N} \frac{\hat{\beta}^2}{\widehat{\text{var}}_R(\hat{\beta})} \sim \chi_1^2. \quad (5.24)$$

Approach 2.

It is well known that a *t*-test is preferred to a normal test when the true variance is estimated. Similarly, we could evaluate a test statistic which follows an *F* distribution rather than a chi-square distribution in order to reduce the effects of confidence interval undercoverage and inflated type I error caused by the use of sandwich estimators. Moreover, according to the equivalence between the *t* and the *F* statistics with numerator degrees of freedom 1, the *F*-test is used instead of the *t*-test here so that test statistics presented here could be extended to more general situations.

Several approaches to determine the denominator degrees of freedom in the *F*-test have been proposed. First, we could consider $N-K$ as the denominator degrees of freedom in the *F*-distribution to lower the inflated test size (Mancl and Derouen, 2001). Therefore, under the hypothesis $H_0 : \beta = 0$ the usual Wald test statistic has the distribution

$$W_{df2} = \frac{\hat{\beta}^2}{\widehat{\text{cov}}_R(\hat{\beta})} \sim F_{1,N-K} \quad (5.25)$$

Approach 3.

Fay and Graubard (2001) proposed an approximate denominator degrees of freedom in the *F*-distribution by taking account of the variability in the sandwich estimator. Extending their approach to ordinal data, the robust Wald test statistic under $H_0 : \beta = 0$ has an *F* distribution

$$W_{df3} = \frac{\hat{\beta}^2}{\widehat{\text{cov}}_R(\hat{\beta})} \sim F_{1,d}\,, \quad (5.26)$$

where the denominator degrees of freedom in (5.26), d is estimated by a function of the variance of the sandwich estimator. Here we give its principle estimating procedures.

Given

$$\hat{\beta}^* = V_M \sum_{i=1}^{2} \sum_{j=1}^{n_i} U_{ij}\,,$$

the numerator in the Wald test statistic (5.34) has the form

$$(C'\hat{\beta}^*)^2 = C'V_M \sum_{i=1}^{2} \sum_{j=1}^{n_i} U_{ij} \sum_{i=1}^{2} \sum_{j=1}^{n_i} U'_{ij} V_M C\,, \quad (5.27)$$

We assume the term $C'V_M U_{ij}$ has mean 0 and the variance ϕ_{ij}^2. Fay and Graubard (2001) showed that

$$\frac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} C'V_M U_{ij} \sum_{i=1}^{2} \sum_{j=1}^{n_i} U'_{ij} V_M C}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} \phi_{ij}^2} \sim \chi_1^2 \quad (5.28)$$

as $N \to \infty$. Substituting (5.27) into (5.28), the numerator in the Wald statistic of (5.26) has an asymptotic chi-square distribution with 1 degree-of-freedom. That is,

$$\frac{(C'\beta^*)^2}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} \phi_{ij}^2} \sim \chi_1^2\,. \quad (5.29)$$

Moreover, Fay and Graubard (2001) showed that the denominator term in the Wald statistic also has an asymptotic chi-square distribution. Let $U = [U_{11}', U_{12}', ..., U_{ij}', ...U_{2n_2}]'$ be a $KN \times 1$ vector, M a $NK \times NK$ block diagonal matrix with *ij*th block equal to $V_M CC' V_M$, $G = I_{KN} - \Omega V_M F'$ assuming I_K an $K \times K$ identity matrix, $F = [I_K, I_K, ..., I_K]'$ a $KN \times K$ matrix, and $\Omega = [D_{11}' V_{11}^{-1} D_{11}, D_{12}' V_{12}^{-1} D_{12}, ...D_{ij}' V_{ij}^{-1} D_{ij}, ..., D'_{2n_2} V_{2n_2}^{-1} D_{2n_2}]'$ a $KN \times K$ matrix. Fay and Graubard (2001) showed that

$$C' \hat{V}_M^{-1} C \approx U' G' MGU$$

and

$$\frac{C' \hat{V}_M^{-1} Cd}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} \phi_i^2} = \frac{U' G' MGUd}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} \phi_i^2} \sim \chi_d^2 \quad (5.30)$$

where

$$d = \frac{\{trace(\Gamma G' MG)\}^2}{trace(\Gamma G' MG \Gamma G' MG)} \quad (5.31)$$

and Γ denotes the estimated covariance matrix with block diagonal ($U_{11}U'_{11}, U_{12}U'_{12}, ..., U_{2n_2}U'_{2n_2}$).

Approach 4.

Alternatively, Pan and Wall (2002) proposed a more general approach to estimate the approximate *F*-distribution by taking account of the variability of the sandwich estimator. Extending their approach to ordinal data, the robust Wald test statistic under $H_0 : \beta = 0$ has an *F* distribution

$$W_{df4} = \frac{\hat{\beta}^2}{\widehat{\mathrm{var}}_R(\hat{\beta})} \sim F_{1,d'}, \quad (5.32)$$

where the denominator degrees of freedom in (5.40), d' is used to distinguish to the denominator degrees of freedom d in (5.26). Next we give principle estimating procedures of d' in (5.32).

Following Pan and Wall (2002), we defined the symbol $\otimes$ as the Kronecker product of two matrices, and $vec(U)$ as an operation which stacks the columns of U below one another. Denote the middle part of V_R as

$$P_{ij} = vec(D'_{ij} V_{ij}^{-1} \mathrm{cov}(Z_{ij}) V_{ij}^{-1} D_{ij}),$$

which has the mean vector

$$\overline{P} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} P_{ij} / N$$

and the empirical covariance estimator

$$\hat{G} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (P_{ij} - \overline{P})(P_{ij} - \overline{P})' / N(N-1).$$

Following Pan and Wall (2002),

$$vec(V_R) = (V_M \otimes V_M) \sum_{i=1}^{2} \sum_{j=1}^{n_i} P_{ij}$$

and its corresponding covariance matrix is

$$\mathrm{cov}(vec(V_R)) = N^2 (V_M \otimes V_M) \hat{G} (V_M \otimes V_M). \quad (5.33)$$

The estimated variance of $\widehat{\mathrm{var}}_R(\hat{\beta})$, denoted as $\hat{\tau}^2$, is the (K^2, K^2) element in $\mathrm{cov}(vec(V_R))$.

Let the mean and variance of $\widehat{\text{var}}_R(\hat{\beta})$ be μ_R and σ_R^2 respectively. Pan and Wall (2002) showed that $\widehat{\text{var}}_R(\hat{\beta})$ has an approximate chi-square distribution

$$\frac{\frac{\mu_R^2}{\sigma_R^2/2}\widehat{\text{var}}_R(\hat{\beta})}{\mu_R} = \frac{2\mu_R\widehat{\text{var}}_R(\hat{\beta})}{\sigma_R^2} \sim \chi^2_{d'}. \quad (5.34)$$

Here $d' = \frac{\mu_R^2}{\sigma_R^2/2}$. On the other hand, under $H_0 : \beta = 0$ we have

$$\frac{\hat{\beta}^2}{\mu_R} \sim \chi_1^2. \quad (5.35)$$

Combining (5.34) and (5.35), Pan and Wall (2002) built an *F* statistic in the form of

$$\frac{\frac{\hat{\beta}^2}{\mu_R}/1}{\frac{2\mu_R\widehat{\text{var}}_R(\hat{\beta})}{\sigma_R^2}/d'} = \frac{\hat{\beta}^2}{\widehat{\text{var}}_R(\hat{\beta})} \sim F_{1,d'} .$$

It is interesting to note that the resulting test statistic is exactly the same as the usual Wald test statistic. Consequently, under $H_0 : \beta = 0$ the Wald test statistic has an approximate *F* distribution,

$$W_{df\,4} = \frac{\hat{\beta}^2}{\widehat{\text{var}}_R(\hat{\beta})} \sim F_{1,d'} . \quad (5.36)$$

where d' is approximated by

$$\hat{d}' = 2\widehat{\text{var}}_R^2(\hat{\beta}) / \hat{\tau}^2 .$$

Approach 5.

Combining degrees-of-freedom adjusted approaches 1 and 2 (i.e., equations 5.24 and 5.25), we derive the Wald test statistic for the hypothesis $H_0 : \beta = 0$ given by

$$W_{df5} = \frac{N-K}{N} \frac{\hat{\beta}^2}{\widetilde{\text{var}}_R(\hat{\beta})} \sim F_{1,N-K} . \qquad (5.37)$$

5.3 Adjustments to the score test

Guo et al. (2005) reported that in contrast to the liberal behaviour of the Wald test, the score test tends to have a smaller test size than the nominal level. They further developed a simple modification to correct this conservative performance of the score test. In this section we adapt their approach to correlated ordinal outcomes.

In the robust score statistic (4.18), both the $\tilde{T}_{(0)}$ term and the $\widetilde{\text{var}}_R(\hat{\beta})$ term are based on the calculation of U_{ij} and this correlation may cause the conservative performance of the robust score test. However, the correlation can be reduced by using the sample variance estimator,

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(U_{ij}-\overline{U})(U_{ij}-\overline{U})',$$

rather than $\sum_{i=1}^{2}\sum_{j=1}^{n_i}U_{ij}U_{ij}'$, to estimate the variance term $\text{cov}(\sum_{i=1}^{2}\sum_{j=1}^{n_i}U_{ij}U_{ij}')$ in $\widetilde{\text{var}}_R(\hat{\beta})$. Let $\overline{U} = \sum_{i=1}^{2}\sum_{j=1}^{n_i}U_{ij} / N$. The sample variance estimator of $\text{cov}(\sum_{i=1}^{2}\sum_{j=1}^{n_i}U_{ij}U_{ij}')$ is given by

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(U_{ij}-\overline{U})(U_{ij}-\overline{U})' \approx \frac{N-1}{N}\sum_{i=1}^{2}\sum_{j=1}^{n_i}U_{ij}U_{ij}' . \qquad (5.38)$$

Consequently, the modified score test statistic for the hypothesis $H_0 : \beta = 0$ is given by

$$S_{BC} = \frac{N}{N-1} \frac{\tilde{T}_{(0)}^2 \widetilde{\text{var}}_M^2(\hat{\beta})}{\widetilde{\text{var}}_R(\hat{\beta})} \sim \chi_1^2 . \qquad (5.39)$$

When the total number of clusters N is large, the modified score test statistic in (5.38) is similar to the regular score statistic in (4.18). When N is small, the factor $N/N-1$ reduces the conservativeness of the regular score statistic.

5.4 Summary

In this chapter, we derived small-sample corrections and modifications of GEE for clustered ordinal outcome data. In particular, we presented five bias-corrected approaches, four degrees-of-freedom-adjusted approaches for the Wald test, and one modified score test. Their performance will be evaluated by simulation in chapter.

Chapter 6

6 Simulation Study: Design

6.1 Introduction

In previous chapters we described statistical approaches for analyses of clustered ordinal outcome data. In this chapter, we outline a simulation study to evaluate the performance of these approaches.

There are two primary objectives. One is to evaluate the accuracy and efficiency of the ANOVA and kappa-type ICC estimators in terms of bias and standard errors. The other is to investigate the 19 test statistics in terms of Type I error and power. These test statistics were presented in the previous chapters, including three direct-adjusted Cochran-Armitage test statistics described in Chapter 3, one GEE model-based, one GEE robust Wald test and one GEE robust score statistic from marginal extensions of proportional odds models, one test statistic from random-effect proportional odds models (i.e., cluster-specific model), one t test statistic from random-effect linear models, and 11 modified GEE test statistics discussed in chapter 5.

The rest of the chapter is organized as follows. A detailed discussion of the parameters used to define the study is given in section 6.2. The methods used to generate the data are presented in section 6.3. Finally, the statistics being evaluated are reviewed in section 6.4. The design of the simulation study follows the guidelines proposed by Burton et al. (2006).

6.2 Parameters used in simulation

The performance of statistical methods considered in this study may depend on the values of the following parameters: the number of clusters per group, cluster sizes, variation in cluster size, the correlation structure among the observations within the same cluster, and proportion of observations falling into each category. These parameters are discussed as follows.

Ordinal outcomes with three categories are most commonly used in health related studies (see Table 1.1). We therefore restrict our attention to clustered ordinal outcome data with three categories.

As discussed in chapter 1, our interest focuses on community intervention trials since most statistical challenges are posed by trials involving a small number of large clusters. For example, we will investigate the performance of the GEE approach when the total number of clusters is less than 40. Furthermore, attention is limited to equal numbers of clusters for each group, with $n_i = n = 10$ and 20 clusters per group. This decision reflects typical practice in cluster randomization trials (e.g., Klar, 1993, p175; Zou, 2002, p64).

Zou et al. (2005) chose 120 as one of the mean cluster sizes for community intervention trials in their study, corresponding to the mean cluster size in trials reported by Murray et al. (1992). In our example data, the mean cluster size is 57. For our simulation study, we selected mean cluster sizes of 50 and 120.

There tends to be considerable variability in cluster size in most cluster randomization trials. The variability of cluster sizes may be measured by an imbalance parameter λ, denoted by

$$\lambda = \frac{1}{1 + CV^2} \qquad (6.1)$$

where CV is the coefficient of variation of the cluster sizes. The parameter λ is equal to one if there is no variability in cluster size and decreases as the imbalance degree in cluster size increases. A value of $\lambda = 0.8$ was found using data from community intervention studies reported by Murray et al. (1992), Villar et al. (2001) , Zou et al. (2005) and the Television School and Family Smoking Prevention and Cessation Project (TVSFP) study (Flay, et al., 1988). Aside from $\lambda = 0.8$, we also included the case $\lambda = 1$ since the ANOVA and kappa-type ICC estimators investigated in this simulation are asymptotically equivalent when the cluster sizes are constant and the number of clusters is large (see section 2.4.3).

Campbell et al. (2005) calculated 220 ANOVA ICCs from 21 datasets and reported a range from 0 to 0.415 with median 0.048. It is known that the ICC value for community intervention trials is much smaller than that for family-type trials. Thus, Hannan et al. (1994) reported that the ICC estimators in his community intervention trial ranged from 0.002 to 0.120 for heart disease risk factors. Donner and Klar (1996) and Zou et al. (2005) selected 0.005 and 0.01 as ICC values representing community type trials in their studies. Correspondingly, we set $\rho = 0.005$ and 0.01 in this simulation. We also included the case $\rho = 0$, where outcome data are independent, as a benchmark against which to evaluate the impact of clustering.

The intervention effect was measured using the log cumulative odds ratio in this study. The cumulative odds ratio θ was chosen as 1.1 and 1.5 for power comparisons as in previous studies (Donner and Donald, 1987; Zhang, 2009). In this simulation study, we consider $\theta = 1.2$ for the power comparisons and $\theta = 1$ for the Type I error comparisons, equivalent to 0.079 and 0 in terms of log odds ratio respectively. In addition, the probabilities of subjects falling into the three categories are close to (0.2, 0.3, 0.5) in the example data (TVSFP). To resemble the example data, we set the expected probabilities in the intervention group, (π_1, π_2, π_3), as (0.2, 0.3, 0.5). As such, the corresponding expected probabilities in the control group are (0.23, 0.31, 0.46) and (0.2, 0.3, 0.5) respectively, corresponding to $\theta = 1.2$ and $\theta = 1$.

There were 48 parameter combinations used in this simulation. For each parameter combination, we generated 1000 independent sets of clustered ordinal data. For each data set, we simulated data for the intervention and control group separately. To reach 1000 replicates, any iteration where iterative procedures failed to converge was replaced by additional data. The full set of parameter values used in this simulation is summarized in Table 6.1.

Table 6.1: Simulation parameters for cluster randomization simulation study

Parameters	Values
Number of ordinal categories (K)	3
Number of clusters per group ($n_1 = n_2 = n$)	10, 20
Mean cluster size (μ)	50, 120
Imbalance degree (λ)	0.8, 1
ICC (ρ)	0, 0.005, 0.01
Probabilities in the intervention group (π_1, π_2, π_3)	(0.2, 0.3, 0.5)
Cumulative odds ratio (θ)	1, 1.2

6.3 Generation of data

6.3.1 Cluster sizes

Three general approaches for generating variable cluster sizes have been used in the literature. The simplest approach is to pre-specify cluster sizes for simulation. For example, in Mancl and Derouen (2001)'s simulation study evaluating their bias-corrected GEE approach, they set 16, 32, or 64 observations in each cluster in order to correspond to a study with 4, 8, or 16 tooth sites in each of the four quadrants of the mouth.

An alternative approach is to assume an empirical distribution of cluster sizes which could be obtained from earlier studies. For instance, Kupper et al. (1986) presented a distribution of cluster sizes (i.e., mouse litters) applicable to dose response modelling in teratologic studies, also used by Ridout et al. (1999) in their simulation study. However, this approach does not allow the average cluster size and the degree of imbalance to be varied.

Donner and Koval (1987) used a more technically sophisticated approach to generate cluster sizes, combining the advantages of the above two approaches. They generated cluster sizes from a negative binomial distribution truncated below one. This approach specifies both the mean cluster size and the degree of imbalance while not restricting the clusters to a few pre-selected sizes. The probability function of the truncated negative binomial distribution generating cluster sizes m_{ij} is given by

$$P(m_{ij}) = \frac{(R+m_{ij}-1)!}{(R-1)!m_{ij}!}(P/Q)^{m_{ij}}(Q^R-1)^{-1}. \quad Q = 1+P \ . \ (6.2)$$

Here the values of (R, P) in (6.2) can be obtained by solving the following nonlinear equations corresponding to the mean and imbalance parameters:

$$\mu = \frac{RPQ^R}{Q^R-1} \qquad (6.3)$$

and

$$\lambda = \frac{\mu}{1+P+RP} \quad (6.4)$$

respectively. Then various cluster sizes m_{ij} can be generated from the truncated negative binomial distribution (6.2) determined by μ and λ.

Since the lower bound of cluster sizes for the truncated negative binomial distribution in equation (6.2) is 1, this approach may yield cluster sizes which are equal or close to 1. However, the minimum cluster size of a community intervention trial is much greater than 1. As such, Zou et al. (2005) used the discrete uniform distribution to generate cluster sizes for community intervention trials. In this simulation study we chose the discrete uniform distribution to generate cluster sizes.

The probability function of the uniform distribution $U(m_l, m_u)$ has the form

$$P(m_{ij}) = \frac{1}{m_u - m_l + 1}, \ m_{ij} = m_l, m_l+1, ..., m_u \quad (6.5)$$

with mean

$$\mu = (m_l + m_u)/2 \qquad (6.6)$$

and variance

$$\sigma^2 = (m_u - m_l)(m_u - m_l + 2)/12 . \qquad (6.7)$$

One can determine the values of (m_l, m_u) by solving equations (6.6) and (6.7) given μ and σ which could be derived by equation (6.1). Then various cluster sizes could be generated from the uniform distribution defined in equation (6.5) determined by the respective mean cluster size and imbalance parameters μ and λ.

The resulting data are then restricted in range with lower bound m_l and upper bound m_u. Therefore, the values of the simulated cluster sizes fall into the suitably chosen range of (m_l, m_u). We list values of (m_l, m_u) in Table 6.2 corresponding to the values of μ and λ presented in Table 6.1. When λ=1, the trial has fixed cluster sizes equal to μ.

Table 6.2: Values of simulation parameters (m,m) corresponding to given (μ,λ)

μ	λ	σ	m_l	m_u
50	0.8	25	7.2	92.8
120	0.8	60	16.6	223.4

6.3.2 Generating clustered ordinal outcome data

The multivariate normal distribution has been widely used to generate correlated categorical data. For example, Jung and Kang (2001) generated clustered ordinal data by generating multivariate normal data with correlation parameter ρ and then discretizing the data using appropriate cut-off values.

In addition, Gange (1995) proposed a procedure for generating multivariate categorical outcomes using an iterative proportional fitting algorithm. However, his approach needs specification of the joint distribution and higher order associations, and requires an iterative procedure to fit the corresponding log-linear models. Also, Biswas (2004) developed algorithms to generate correlated ordinal outcomes for some specific correlation structures. However, generalizations of his algorithms to other correlation

structures are doubtful. Moreover, the algorithm itself lacks practical meaning, and Biswas (2004) did not provide evaluations for his approach using simulation.

Demirtas (2006) proposed a method for generating multivariate ordinal outcomes with specified marginal distribution and correlation structure. His method relies on simulating correlated binary outcomes as an intermediate step and then converting them to correlated ordinal outcomes. However, it is computationally burdensome as it requires iterative procedures to compute the proper correlations for binary data.

Correlated ordinal outcome data may also be generated from cluster-specific models. However, the odds ratios from a mixed effects model are larger on average as they are estimating a different (larger) parameter compared to those from marginal models (e.g., Ten Have et al., 1996). Since we are primarily interested in examining the statistical properties of marginal models (e.g., GEE), a marginal model was used to simulate clustered ordinal outcome data in this study.

In particular, correlated binary data may be generated from the beta-binomial distribution (e.g., Donner and Klar, 1996; Donner et al., 1994; Bellamy et al., 2000). This simple method could be extended to generate correlated ordinal data by using a dirichlet-multinomial distribution (e.g., Tsou and Shen, 2008; Lui et al., 1999). This simplicity can be obtained since attention here has been limited to trials where responses of all cluster members are assumed to be equally correlated with a single cluster-level covariate (i.e., intervention vs. control). Therefore, we simulated clustered ordinal outcome data from a Dirichlet-multinomial distribution in this study.

In the remainder of this section we describe the Dirichlet-multinomial distribution and its use in generating correlated ordinal outcomes.

Let $Y_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})$ be the vector of counts for the *ij*th cluster, where Y_{ijk} denotes the number of subjects in the *ij*th cluster falling into the *k*th category. Let $P_{ij} = (P_{ij1}, P_{ij2}, P_{ij3})$, where P_{ijk} is the probability of a subject in the *ij*th cluster falling into the *k*th category. To

account for the variation between clusters, we further assume that $P_{ij} = (P_{ij1}, P_{ij2}, P_{ij3})$ are from a Dirichlet distribution of the form

$$\frac{\Gamma(\theta_1 + \theta_2 + \theta_3)}{\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)} P_{ij1}^{\theta_1 - 1} P_{ij2}^{\theta_2 - 1} P_{ij3}^{\theta_3 - 1}, \qquad (6.8)$$

where $\theta_k > 0$. Then given P_{ij}, ordinal outcomes $Y_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})$ have a trinomial distribution with parameters m_{ij} and P_{ij}. Consequently, the resulting data has a Dirichlet-trinomial distribution of the form

$$P(Y_{ij1}, Y_{ij2}, Y_{ij3}) = \frac{m_{ij}!\Gamma(\theta_1 + \theta_2 + \theta_3)\Gamma(Y_{ij1} + \theta_1)\Gamma(Y_{ij2} + \theta_2)\Gamma(Y_{ij3} + \theta_3)}{Y_{ij1}!Y_{ij2}!Y_{ij3}!\Gamma(m_{ij} + \theta_1 + \theta_2 + \theta_3)\Gamma(\theta_1)\Gamma(\theta_2)\Gamma(\theta_3)}. \qquad (6.9)$$

In equation (6.8), the mean of P_{ijk} is given by

$$E(P_{ijk}) = \pi_k = \theta_k / (\theta_1 + \theta_2 + \theta_3) \qquad (6.10)$$

and the constant ICC is of the form

$$\rho = 1/(1 + \theta_1 + \theta_2 + \theta_3). \qquad (6.11)$$

Therefore, we can determine the values of $(\theta_1, \theta_2, \theta_3)$ given the values of π_k and ρ by solving equations (6.10) and (6.11).

In summary, the data generation in this study may be described by the following steps:

1. Set up the values of the mean cluster size μ and imbalance parameter λ;
2. Generate various cluster sizes m_{ij} from the uniform distribution (6.5);
3. Given (π_1, π_2, π_3) and ρ, calculate values of $(\theta_1, \theta_2, \theta_3)$ through equations (6.10) and (6.11);
4. Generate proportions $(P_{ij1}, P_{ij2}, P_{ij3})$ from a Dirichlet distribution in (6.8) with parameters $(\theta_1, \theta_2, \theta_3)$ for each cluster;

5. Generate correlated ordinal outcomes $(Y_{ij1}, Y_{ij2}, Y_{ij3})$ from a multinomial distribution with parameters m_{ij} and $(P_{ij1}, P_{ij2}, P_{ij3})$ for each cluster.

6.4 Evaluation measures

The evaluation measures used to compare the performance of the proposed approaches were computed as following:

1) Average value of estimates;

2) Relative bias computed as the deviation of the average observed value from the true value (parameter) divided by the true value . Positive relative bias represents an overestimate of the parameter and negative bias represents an underestimate of the parameter. The relative bias is not applicable for the ICC estimate when its true value is set as zero.

3) The standard errors of estimated regression coefficients (log odds ratio) were computed as the empirical standard error of 1000 estimate values. Note that the standard error of estimated regression coefficients (log odds ratio) from GEE extensions of ordinal logistic regression is obtained from the sandwich estimator.

4) The type I error rate was calculated as the proportion of simulation samples generated under the null hypothesis which have p-values less than or equal to the nominal 5% significance level.

5) The statistical power was calculated as the proportion of simulation samples generated under the alternative hypothesis which have p-values less than or equal to

the nominal 5% significance level, given that the corresponding test statistic provides a valid type I error rate.

Since 1000 iterations were used, the approximate 95% confidence interval for a five percent rejection rate is (0.031, 0.069). Therefore, the statistical test is overly conservative for type I error rates less than 0.03, and overly liberal for type I error rates greater than 0.07 (Bradley, 1978). Power comparisons were limited to those test statistics with valid type I error rates.

6.4.1 Investigation of the ICC estimators

In this objective, we evaluated properties of the ANOVA and kappa-type ICC estimator. Both spaced scores (i.e.,$1,2,3$) and midrank scores were considered in calculating the ICC estimators, as summarized in Table 6.3. The subscript 'M' denotes the ICC estimator calculated by using midranks. In addition, since Cohen's regular kappa $\hat{\kappa}$ (Cohen, 1960) is also a frequently used statistic measuring the agreement of ordinal outcomes, we compared properties of $\hat{\rho}_{\kappa}$ and $\hat{\rho}_{A}$ to $\hat{\kappa}$ as well. The ICC estimators, defined in Chapter 2, are compared were listed in Table 6.3 in terms of the average, relative bias, range, and standard error.

Since negative ICCs are generally considered implausible in the context of cluster randomization trials, negative ICC estimators are usually set to zero. In the present study, we also followed this practice.

Table 6.3: ICC estimators evaluated in simulation study

ICC estimator	With scores 1,2,3	With midrank scores
ANOVA ICC estimator	$\hat{\rho}_{A}$	$\hat{\rho}_{A(M)}$
Kappa-type ICC estimator	$\hat{\rho}_{\kappa}$	$\hat{\rho}_{k(M)}$

6.4.2 Evaluation of test statistics

Adjusted Cochran-Armitage tests

We evaluated the performance of three direct-adjusted Cochran-Armitage test statistics for clustered ordinal outcome data. Both $\hat{\rho}_{\kappa}$ and $\hat{\rho}_{A}$ were used as estimates of the ICC in the test statistics. In addition, since both equally-spaced score and midrank scores were considered for $\hat{\rho}_{\kappa}$ and $\hat{\rho}_{A}$, adjusted Cochran-Armitage test statistics were also calculated using these two scoring schemes respectively. The twelve test statistics evaluated under this objective are listed in Table 6.4, and compared in terms of Type I errors and statistical power.

As reviewed in section 1.5.1, Jung and Kang (2001) proposed another simple adjustment to the Cochran-Armitage test for clustered ordinal outcomes. We also compared adjusted Cochran-Armitage tests to Jung and Kang (2001)'s approach (i.e., χ^2_J and $\chi^2_{J(M)}$).

Table 6.4: Adjusted Cochran-Armitage test statistics evaluated in simulation

	With $\hat{\rho}_{A}$	With $\hat{\rho}_{\kappa}$	With $\hat{\rho}_{A(M)}$	With $\hat{\rho}_{k(M)}$
Adjusted Cochran-Armitage test (I)	χ^2_{A1}	$\chi^2_{\kappa 1}$	$\chi^2_{A1(M)}$	$\chi^2_{\kappa 1(M)}$
Adjusted Cochran-Armitage test (II)	χ^2_{A2}	$\chi^2_{\kappa 2}$	$\chi^2_{A2(M)}$	$\chi^2_{\kappa 2(M)}$
WLS Cochran-Armitage test	χ^2_{A3}	$\chi^2_{\kappa 3}$	$\chi^2_{A3(M)}$	$\chi^2_{\kappa 3(M)}$

Comparisons of model-based approaches

Test statistics from marginal extensions of proportional odds regression models and mixed-effect proportional odds regression models (i.e., cluster-specific models) were compared in terms of type I error rates and power. In particular, marginal models were fitted by the GEE approach using an independent working correlation. Since mixed effects linear models are commonly used to fit clustered ordinal outcomes, we also compared marginal and cluster-specific models to random effects linear models, where the test statistic is the t-statistic expressed as the ratio of the parameter estimate to its standard error. The SAS procedures used in the above marginal, cluster-specific, and random effects linear models were PROC GENMOD, PROC NLMIXED, and PROC GLIMMIX respectively. The test statistics under this objective are listed as the first four

statistics in Table 6.5, denoted as W_M , W_R , χ^2_{CS} , and T_{Linear} correspondingly. They were previously defined in Chapter 4.

Table 6.5: Model-based test statistics evaluated in simulation study

Test statistic	Test name
W_M	Model-based Wald test statistic
W_R	Robust Wald test statistics
S_R	Robust score test statistic
χ^2_{CS}	Chi-square test statistic from cluster-specific models
T_{Linear}	T-test statistic from random effects linear model
W_{BC1}	Bias-corrected robust Wald test: Approach 1
W_{BC2}	Bias-corrected robust Wald test: Approach 2
W_{BC3}	Bias-corrected robust Wald test: Approach 3
W_{BC4}	Bias-corrected robust Wald test: Approach 4
W_{BC5}	Bias-corrected robust Wald test: Approach 5
W_{df1}	Degrees-of-freedom-adjusted Wald test: Approach 1
W_{df2}	Degrees-of-freedom-adjusted Wald test: Approach 2
W_{df3}	Degrees-of-freedom-adjusted Wald test: Approach 3
W_{df4}	Degrees-of-freedom-adjusted Wald test: Approach 4
W_{df5}	Degrees-of-freedom-adjusted Wald test: Approach 5
S_{BC}	Modified robust score test

In particular, we give special attention to the magnitudes of the regression coefficient estimate $\hat{\beta}$ and its standard error in the marginal and cluster-specific models (Ten Have et al., 1996), as discussed in section 4.4. Note that the standard error of estimated regression coefficients (i.e., log odds ratios) from GEE extensions of ordinal logistic regression is obtained from the sandwich estimator. The estimates are listed in Table 6.6.

Table 6.6: Regression coefficient estimates and their standard errors from marginal and cluster-specific models

Approaches	Parameter estimate	Standard errors
Marginal models	$\hat{\beta}_{GEE}$	$SE(\hat{\beta}_{GEE})$
Cluster-specific models	$\hat{\beta}_{CS}$	$SE(\hat{\beta}_{CS})$

Evaluation of small-sample adjustments to GEE

Five bias-corrected and four degrees-of-freedom-adjusted approaches for the robust Wald test and one correction approach for the robust score test discussed in Chapter 5 were investigated. Therefore, including four model-based test statistics discussed previously (i.e., W_M , W_R ,W_{CS} , and T_{Linear}), a total of 16 model-based test statistics were compared and summarized in Table 6.5. They were previously defined in Chapter 5.

6.5 Computation implementation

All the computer programs for the simulation study were written in SAS V.9.2 (SAS Institute, Inc, Cary, NC) and run on a PC Workstation. Specifically, the methods of GEE, cluster-specific models, and random effects models were carried out with SAS procedures PROC GENMOD, PROC NLMIXED, and PROC GLIMMIX and correspondingly.

Chapter 7

7 Simulation Results

7.1 Introduction

In Chapter 6 we described the design of the simulation study which was used to investigate and compare statistical approaches presented in earlier chapters. In this chapter, the results of this study are presented and tabulated in the order of objectives outlined in Section 6.4.

In particular, there are five sections in this chapter. Section 7.2 compares the ANOVA, kappa-type ICC estimators, and Cohen's (1960) regular kappa estimates in terms of relative bias and standard errors. Section 7.3 discusses the type I error rate and power of the adjusted Cochran-Armitage tests. The modelling tests from marginal extensions of proportional odds ratio models and mixed-effects ordinal regression models (i.e., cluster-specific models) and their adjustments are discussed in Section 7.4. The bias of estimated regression coefficients as obtained from marginal and cluster-specific models and their standard errors are summarized in section 7.5.

In previous studies, attention was given to convergence problems for the iterative procedures. However, there was no problem reaching convergence when running the computer programs in this simulation study, probably because the cluster sizes were large, i.e., 50 and 120. This conclusion is consistent with Zhang's (2009) results. In her simulation study, the SAS procedure PROC GLIMMIX was used to fit the cluster-specific models for clustered binary outcome data, and convergence problems only occurred when generating data in which the cluster size is 15, i.e., the smallest size used in the study.

7.2 Estimation of Intracluster Correlation Coefficients

Simulation results for the regular kappa estimator, two ANOVA ICC estimators, and two kappa-type ICC estimators are displayed in Table 7.1 through Table 7.12. The parameters of interest include the cumulative odds ratio θ, number of clusters from each group n,

mean cluster size μ, imbalance degree for cluster size λ and intracluster correlation coefficient ρ. Each table displays the results for ICC estimators for each parameter combination.

Overall, Cohen (1960)'s regular kappa estimator $\hat{k}$ had the least number of negative values compared to the ANOVA estimators ($\hat{\rho}_A$ and $\hat{\rho}_{A(M)}$) and kappa-type ICC estimators ($\hat{\rho}_k$ and $\hat{\rho}_{k(M)}$). In addition, all five estimators had less negative values for fixed clustered sizes than for variable cluster sizes.

When cluster sizes were fixed (shown in Table 7.1 through Table 7.6), all four estimators $\hat{\rho}_A, \hat{\rho}_{A(M)}$, $\hat{\rho}_k$ and $\hat{\rho}_{k(M)}$ had a similar number of negative estimates. However, when cluster sizes were variable (shown in Table 7.7 through Table 7.12), $\hat{\rho}_A$ and $\hat{\rho}_{A(M)}$ had a smaller number of negative estimators than $\hat{\rho}_k$ and $\hat{\rho}_{k(M)}$, with the difference between the numbers of negative values becoming larger when the cluster size μ or the number of clusters n increased. In addition, using midrank scores led to slightly less negative kappa-type estimators.

In addition, negative ICC estimators are set to zero since negative ICCs are generally considered implausible in the context of cluster randomization trials. This practice may elevate the resulted average of ICC estimates.

Simulation results after truncating the negative ICC estimators are displayed in the last two columns in Table 7.1 through Table 7.12. Overall, all ICC estimators were closer to the true values when either cluster sizes or the number of clusters become large. To be more specific, kappa-type estimators were more close to the true values than ANOVA estimators when cluster sizes were fixed and small. Conversely, ANOVA ICCs had relatively smaller bias in the case of variable cluster sizes. In addition, using midrank scores yielded less biased estimators for kappa-type ICCs in the case of variable cluster sizes. The standard errors of the all ICC estimators (not shown) were approximately zero to three decimal places for most parameter combinations.

7.3 Adjusted Cochran-Armitage Tests

In Chapter 3, we discussed three adjusted C-A tests for analysis of clustered ordinal data. In the case of fixed cluster sizes, all three test statistics provided equivalent results. We therefore only listed one simulation result for all three statistics. In contrast, for variable cluster sizes, we listed results for each of the three adjusted C-A statistics separately.

As discussed in Section 6.2.4, for nominal level 0.05 the empirical rate will be regarded as satisfactory provided it lies in the range (3.1, 6.9)%. Overall power is discussed only for those tests which have acceptable Type I error.

7.3.1 Type I Error rates

Estimated type I error rates in the case of fixed cluster sizes presented in Table 7.13 show that all adjusted C-A test statistics maintain the nominal significance level of 5% reasonably well, with the type I error rates for variable cluster sizes listed in Table 7.14. The observed significance levels of 0.069 or higher and 0.031 or lower are highlighted. Overall, the three adjusted C-A tests produced similar type I error rates when using the same ICC estimator. However, when $\rho = 0.01$, the C-A tests using kappa-type estimators resulted in liberal type I error rates. Inversely, when $\rho = 0$ and $\mu = 120$, the C-A tests using kappa-type estimators resulted in conservative type I error rates. In addition, tests using the ANOVA ICC estimators produced inflated type I errors when m = 50, n = 10 and $\rho = 0.01$.

7.3.2 Power

Tables 7.15 and 7.16 present the statistical power results for the three adjusted C-A tests in the case of fixed cluster sizes and variable cluster sizes respectively. The evaluation of power is only sensible when the corresponding test statistic has a valid Type I error rate. Thus, the adjusted C-A tests at parameter combinations which showed liberal or conservative type I error rates were excluded.

Overall, the power of adjusted C-A tests was consistently larger for data with fixed cluster sizes than for data with variable cluster sizes. Also, the power of the adjusted C-A

tests increased when cluster sizes and cluster numbers increased. Inversely, the power tended to decrease when the magnitude of the ICC increased.

In particular, when cluster sizes are fixed, adjusted C-A tests using kappa-type ICC estimates have the greatest power. However, when cluster sizes are variable, the WLS C-A tests (i.e., $\chi^2_{A3(M)}$) using ANOVA ICC estimates with midrank scores have the greatest power among the tests which are valid.

7.4 Model-based Methods

7.4.1 Type I Error Rates

Tables 7.17 and 7.18 present the observed type I error rates of the modelling tests for fixed and variable cluster sizes respectively. Overall, the GEE model-based test W_M and GEE robust Wald test W_R tend to result in a liberal type I error rate, especially when the number of clusters is equal to 10. As expected, the liberal behaviors of GEE robust Wald tests were consistently improved by all GEE adjusted methods.

For fixed cluster sizes, type I error rates for all adjusted methods maintained the nominal level for all parameter combinations. However, for variable cluster sizes, the rejection rates of the adjusted methods W_{BC2}, W_{BC3}, W_{df1}, and W_{df2} were still high under most parameter combinations with $n=10$. In contrast, W_{df3} and W_{df4} overcorrected the GEE robust Wald test and resulted in overly conservative type I errors under most parameter combinations. The adjustment methods W_{BC1}, W_{BC4} and W_{df5} showed fairly unbiased type I error rates at all parameter combinations investigated.

The type I error rates of robust score tests were valid under all parameter combinations. The adjusted score tests tended to elevate them as expected.

7.4.2 Power

Tables 7.19 and 7.20 present the power of the modelling tests for fixed and variable cluster sizes respectively. Overall, the power of all tests tended to decrease when the

magnitude of the true ICCs increased. Conversely, the power tended to increase when the number of clusters and the cluster sizes became large.

For fixed cluster sizes, the adjusted method W_{BC2} yielded the highest statistical power among the methods that were valid. For variable cluster sizes, however, the power of W_{BC1} was the greatest among the methods that were valid.

7.5 Relationship between Marginal and Cluster-specific Models

Table 7.21 and 7.22 show the estimated regression coefficients of the marginal and cluster-specific models and their corresponding standard errors. Under most parameter combinations, as expected, the absolute values of the marginal model coefficient estimates are smaller than the cluster-specific model coefficient estimates. The standard errors of estimates from the marginal models are also smaller than those from the cluster-specific models. In particular, the discrepancy between estimates from the two models tends to increase when the fixed cluster size becomes larger. However, this trend does not hold for variable cluster sizes.

Table 7.1: Properties of ICC estimators: based on 1000 simulations of trials with n clusters of size μ per group, cumulative odds ratio $\theta = 1$, intracluster correlation $\rho = 0$, and fixed cluster size $\lambda = 1$

Parameters			Descriptive statistics				Descriptive statistics (setting negative estimates to zero)
μ	n	ρ [1]	Average (× 100)	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)
50	10	$\hat{\kappa}$	-0.10	-1.21	1.66	0.59	0.14
		$\hat{\rho}_A$	0.00	-1.52	2.38	0.54	0.27
		$\hat{\rho}_{A(M)}$	0.00	-1.5	2.36	0.53	0.27
		$\hat{\rho}_{\kappa}$	0.00	-1.37	2.15	0.54	0.24
		$\hat{\rho}_{\kappa(M)}$	0.00	-1.35	2.18	0.53	0.24
	20	$\hat{\kappa}$	-0.04	-0.91	1.11	0.56	0.11
		$\hat{\rho}_A$	0.02	-1.13	1.74	0.52	0.20
		$\hat{\rho}_{A(M)}$	0.03	-1.14	1.73	0.52	0.20
		$\hat{\rho}_{\kappa}$	0.02	-1.07	1.66	0.52	0.19
		$\hat{\rho}_{\kappa(M)}$	0.02	-1.08	1.65	0.52	0.19
120	10	$\hat{\kappa}$	-0.04	-0.49	0.66	0.62	0.06
		$\hat{\rho}_A$	0.00	-0.61	1.13	0.55	0.11
		$\hat{\rho}_{A(M)}$	0.00	-0.61	0.99	0.53	0.11
		$\hat{\rho}_{\kappa}$	0.00	-0.55	1.02	0.55	0.10
		$\hat{\rho}_{\kappa(M)}$	0.00	-0.55	0.89	0.54	0.10
	20	$\hat{\kappa}$	-0.02	-0.36	0.45	0.57	0.05
		$\hat{\rho}_A$	0.00	-0.5	0.97	0.55	0.08
		$\hat{\rho}_{A(M)}$	0.00	-0.47	0.91	0.54	0.08
		$\hat{\rho}_{\kappa}$	0.00	-0.48	0.92	0.55	0.07
		$\hat{\rho}_{\kappa(M)}$	0.00	-0.44	0.86	0.54	0.07

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.2: Properties of ICC estimators: based on 1000 simulations of trials with n clusters of size μ per group, cumulative odds ratio $\theta = 1.2$, intracluster correlation $\rho = 0$, and fixed cluster size $\lambda = 1$

Parameters			Descriptive statistics				Descriptive statistics (setting negative estimates to zero)
μ	n	ρ [1]	Average (× 100)	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)
50	10	$\hat{\kappa}$	0.04	-1.15	2.26	0.49	0.21
		$\hat{\rho}_A$	-0.02	-1.51	2.79	0.55	0.26
		$\hat{\rho}_{A(M)}$	-0.01	-1.43	2.87	0.56	0.26
		$\hat{\rho}_{\kappa}$	-0.02	-1.36	2.52	0.55	0.23
		$\hat{\rho}_{\kappa(M)}$	-0.01	-1.29	2.62	0.56	0.23
	20	$\hat{\kappa}$	0.06	-0.96	1.61	0.45	0.16
		$\hat{\rho}_A$	-0.03	-1.25	1.99	0.56	0.16
		$\hat{\rho}_{A(M)}$	-0.03	-1.13	1.89	0.57	0.16
		$\hat{\rho}_{\kappa}$	-0.03	-1.19	1.9	0.56	0.16
		$\hat{\rho}_{\kappa(M)}$	-0.03	-1.07	1.81	0.57	0.16
120	10	$\hat{\kappa}$	0.10	-0.44	1.17	0.37	0.14
		$\hat{\rho}_A$	0.02	-0.57	1.52	0.52	0.12
		$\hat{\rho}_{A(M)}$	0.02	-0.58	1.49	0.53	0.12
		$\hat{\rho}_{\kappa}$	0.01	-0.51	1.37	0.52	0.11
		$\hat{\rho}_{\kappa(M)}$	0.01	-0.52	1.37	0.53	0.11
	20	$\hat{\kappa}$	0.11	-0.27	0.66	0.24	0.13
		$\hat{\rho}_A$	-0.01	-0.49	0.77	0.53	0.07
		$\hat{\rho}_{A(M)}$	-0.01	-0.49	0.73	0.54	0.07
		$\hat{\rho}_{\kappa}$	-0.01	-0.46	0.73	0.53	0.07
		$\hat{\rho}_{\kappa(M)}$	-0.01	-0.47	0.69	0.54	0.07

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.3: Properties of ICC estimators: based on 1000 simulations of trials with n clusters of size μ per group, cumulative odds ratio $\theta = 1$, intracluster correlation $\rho = 0.005$, and fixed cluster size $\lambda = 1$

Parameters			Descriptive statistics					Descriptive statistics (setting negative estimates to zero)	
μ	n	ρ [1]	Average (× 100)	Relative bias	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)	Relative bias
50	10	$\hat{\kappa}$	0.39	-0.22	-1.09	2.51	0.27	0.47	-0.07
		$\hat{\rho}_A$	0.54	0.08	-1.54	4.34	0.29	0.66	0.31
		$\hat{\rho}_{A(M)}$	0.54	0.09	-1.66	4.26	0.29	0.66	0.32
		$\hat{\rho}_{\kappa}$	0.49	-0.03	-1.39	3.93	0.29	0.59	0.18
		$\hat{\rho}_{\kappa(M)}$	0.49	-0.01	-1.49	3.83	0.29	0.60	0.20
	20	$\hat{\kappa}$	0.46	-0.09	-0.53	2.17	0.12	0.48	-0.05
		$\hat{\rho}_A$	0.50	0.01	-0.78	2.87	0.20	0.55	0.11
		$\hat{\rho}_{A(M)}$	0.51	0.02	-0.74	2.73	0.20	0.56	0.12
		$\hat{\rho}_{\kappa}$	0.48	-0.04	-0.75	2.73	0.20	0.53	0.06
		$\hat{\rho}_{\kappa(M)}$	0.49	-0.03	-0.69	2.61	0.20	0.53	0.07
120	10	$\hat{\kappa}$	0.43	-0.14	-0.35	1.34	0.01	0.44	-0.13
		$\hat{\rho}_A$	0.49	-0.03	-0.50	2.22	0.12	0.51	0.01
		$\hat{\rho}_{A(M)}$	0.48	-0.03	-0.53	2.28	0.13	0.50	0.01
		$\hat{\rho}_{\kappa}$	0.44	-0.12	-0.45	2.00	0.12	0.46	-0.09
		$\hat{\rho}_{\kappa(M)}$	0.44	-0.13	-0.48	2.07	0.13	0.45	-0.09
	20	$\hat{\kappa}$	0.46	-0.08	-0.13	1.21	0.01	0.46	-0.08
		$\hat{\rho}_A$	0.50	0.00	-0.26	1.63	0.03	0.50	0.00
		$\hat{\rho}_{A(M)}$	0.50	0.00	-0.23	1.57	0.03	0.50	0.00
		$\hat{\rho}_{\kappa}$	0.47	-0.05	-0.25	1.55	0.03	0.48	-0.05
		$\hat{\rho}_{\kappa(M)}$	0.47	-0.05	-0.22	1.49	0.03	0.48	-0.05

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.4: Properties of ICC estimators: based on 1000 simulations of trials with *n* clusters of size μ per group, cumulative odds ratio ϑ = 1.2, intracluster correlation ρ = 0.005, and fixed cluster size λ = 1

Parameters			Descriptive statistics					Descriptive statistics (setting negative estimates to zero)	
μ	n	ρ [1]	Average (× 100)	Relative bias	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)	Relative bias
50	10	$\hat{\kappa}$	0.52	0.05	-0.93	2.69	0.20	0.57	0.15
		$\hat{\rho}_A$	0.52	0.04	-1.54	3.82	0.27	0.63	0.26
		$\hat{\rho}_{A(M)}$	0.52	0.04	-1.59	3.98	0.28	0.63	0.26
		$\hat{\rho}_{\kappa}$	0.47	-0.06	-1.38	3.45	0.27	0.57	0.14
		$\hat{\rho}_{\kappa(M)}$	0.47	-0.06	-1.43	3.60	0.28	0.57	0.14
	20	$\hat{\kappa}$	0.56	0.12	-0.59	2.15	0.07	0.57	0.14
		$\hat{\rho}_A$	0.49	-0.02	-1.17	2.57	0.19	0.54	0.08
		$\hat{\rho}_{A(M)}$	0.49	-0.03	-1.08	2.55	0.19	0.54	0.07
		$\hat{\rho}_{\kappa}$	0.46	-0.07	-1.11	2.45	0.19	0.51	0.02
		$\hat{\rho}_{\kappa(M)}$	0.46	-0.07	-1.03	2.40	0.19	0.51	0.02
120	10	$\hat{\kappa}$	0.54	0.08	-0.17	1.82	0.02	0.54	0.09
		$\hat{\rho}_A$	0.48	-0.04	-0.46	2.43	0.13	0.50	-0.01
		$\hat{\rho}_{A(M)}$	0.48	-0.03	-0.48	2.53	0.13	0.50	0.00
		$\hat{\rho}_{\kappa}$	0.43	-0.14	-0.41	2.20	0.13	0.45	-0.10
		$\hat{\rho}_{\kappa(M)}$	0.44	-0.13	-0.43	2.29	0.12	0.45	-0.10
	20	$\hat{\kappa}$	0.60	0.19	-0.08	1.51	0.00	0.60	0.19
		$\hat{\rho}_A$	0.50	0.01	-0.39	1.56	0.04	0.51	0.01
		$\hat{\rho}_{A(M)}$	0.50	0.00	-0.42	1.59	0.04	0.51	0.01
		$\hat{\rho}_{\kappa}$	0.48	-0.04	-0.37	1.48	0.04	0.48	-0.04
		$\hat{\rho}_{\kappa(M)}$	0.48	-0.05	-0.40	1.51	0.04	0.48	-0.04

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.5: Properties of ICC estimators: based on 1000 simulations of trials with n clusters of size μ per group, cumulative odds ratio $\theta = 1$, intracluster correlation $\rho = 0.01$, and fixed cluster size $\lambda = 1$

Parameters			Descriptive statistics of ICC estimators					Descriptive statistics of ICC estimators after truncating negative estimates	
μ	n	ρ [1]	Average (× 100)	Relative bias	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)	Relative bias
50	10	$\hat{\kappa}$	0.86	-0.14	-0.69	3.42	0.09	0.88	-0.12
		$\hat{\rho}_A$	1.04	0.04	-1.50	5.46	0.14	1.09	0.09
		$\hat{\rho}_{A(M)}$	1.03	0.03	-1.35	6.14	0.14	1.08	0.08
		$\hat{\rho}_{\kappa}$	0.94	-0.06	-1.35	4.95	0.14	0.98	-0.02
		$\hat{\rho}_{\kappa(M)}$	0.93	-0.07	-1.23	5.63	0.14	0.98	-0.02
	20	$\hat{\kappa}$	0.94	-0.06	-0.34	2.50	0.02	0.95	-0.05
		$\hat{\rho}_A$	1.03	0.03	-0.48	3.49	0.05	1.04	0.04
		$\hat{\rho}_{A(M)}$	1.03	0.03	-0.55	3.64	0.05	1.04	0.04
		$\hat{\rho}_{\kappa}$	0.98	-0.02	-0.45	3.32	0.05	0.99	-0.01
		$\hat{\rho}_{\kappa(M)}$	0.98	-0.02	-0.52	3.40	0.05	0.99	-0.01
120	10	$\hat{\kappa}$	0.91	-0.09	-0.05	2.30	0.00	0.91	-0.09
		$\hat{\rho}_A$	1.00	0.00	-0.32	3.61	0.02	1.01	0.01
		$\hat{\rho}_{A(M)}$	1.01	0.01	-0.39	3.63	0.02	1.01	0.01
		$\hat{\rho}_{\kappa}$	0.91	-0.09	-0.28	3.26	0.02	0.91	-0.09
		$\hat{\rho}_{\kappa(M)}$	0.91	-0.09	-0.35	3.28	0.02	0.91	-0.09
	20	$\hat{\kappa}$	0.95	-0.05	0.14	2.20	0.00	0.95	-0.05
		$\hat{\rho}_A$	0.98	-0.02	-0.15	2.60	0.00	0.98	-0.02
		$\hat{\rho}_{A(M)}$	0.98	-0.02	-0.09	2.49	0.00	0.98	-0.02
		$\hat{\rho}_{\kappa}$	0.93	-0.07	-0.14	2.47	0.00	0.93	-0.07
		$\hat{\rho}_{\kappa(M)}$	0.94	-0.06	-0.08	2.37	0.00	0.94	-0.06

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.6: Properties of ICC estimators: based on 1000 simulations of trials with *n* clusters of size μ per group, cumulative odds ratio θ = 1.2, intracluster correlation ρ = 0.01, and fixed cluster size λ = 1

Parameters			Descriptive statistics					Descriptive statistics (setting negative estimates to zero)	
μ	n	ρ [1]	Average (× 100)	Relative bias	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)	Relative bias
50	10	$\hat{\kappa}$	0.99	-0.01	-0.64	3.70	0.06	1.01	0.01
		$\hat{\rho}_A$	1.01	0.01	-1.22	4.79	0.16	1.06	0.06
		$\hat{\rho}_{A(M)}$	1.01	0.01	-1.12	4.51	0.16	1.06	0.06
		$\hat{\rho}_{\kappa}$	0.91	-0.09	-1.09	4.34	0.16	0.96	-0.04
		$\hat{\rho}_{\kappa(M)}$	0.91	-0.09	-0.99	4.09	0.16	0.96	-0.04
	20	$\hat{\kappa}$	1.05	0.05	-0.33	3.37	0.01	1.05	0.05
		$\hat{\rho}_A$	0.98	-0.02	-0.80	3.23	0.06	0.99	-0.01
		$\hat{\rho}_{A(M)}$	0.98	-0.02	-0.90	3.48	0.06	1.00	0.00
		$\hat{\rho}_{\kappa}$	0.93	-0.07	-0.76	3.08	0.06	0.95	-0.05
		$\hat{\rho}_{\kappa(M)}$	0.94	-0.06	-0.84	3.31	0.06	0.95	-0.05
120	10	$\hat{\kappa}$	1.02	0.02	-0.14	2.82	0.00	1.02	0.02
		$\hat{\rho}_A$	0.93	-0.07	-0.37	4.13	0.03	0.94	-0.06
		$\hat{\rho}_{A(M)}$	0.94	-0.06	-0.39	4.02	0.03	0.94	-0.06
		$\hat{\rho}_{\kappa}$	0.84	-0.16	-0.33	3.73	0.03	0.85	-0.15
		$\hat{\rho}_{\kappa(M)}$	0.85	-0.15	-0.35	3.60	0.03	0.85	-0.15
	20	$\hat{\kappa}$	1.10	0.10	0.25	2.28	0.00	1.10	0.10
		$\hat{\rho}_A$	1.01	0.01	-0.01	2.48	0.00	1.01	0.01
		$\hat{\rho}_{A(M)}$	1.01	0.01	0.00	2.44	0.00	1.01	0.01
		$\hat{\rho}_{\kappa}$	0.96	-0.04	-0.01	2.36	0.00	0.96	-0.04
		$\hat{\rho}_{\kappa(M)}$	0.96	-0.04	0.00	2.33	0.00	0.96	-0.04

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.7: Properties of ICC estimators: based on 1000 simulations of trials with *n* clusters of size μ per group, cumulative odds ratio $\theta = 1$, intracluster correlation $\rho = 0$, and variable cluster size $\lambda = 0.8$

Parameters			Descriptive statistics					Descriptive statistics (setting negative estimates to zero)	
μ	n	ρ [1]	Average (× 100)	Relative bias	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)	Relative bias
50	10	$\hat{\kappa}$	-0.12	-1.25	-1.41	1.48	0.63	0.13	-0.73
		$\hat{\rho}_A$	-0.04	-1.07	-1.72	2.76	0.56	0.25	-0.50
		$\hat{\rho}_{A(M)}$	-0.08	-1.16	-1.74	2.54	0.58	0.23	-0.54
		$\hat{\rho}_{\kappa}$	-0.06	-1.12	-7.83	8.01	0.53	0.90	0.79
		$\hat{\rho}_{\kappa(M)}$	-0.04	-1.08	-6.33	6.07	0.53	0.74	0.48
	20	$\hat{\kappa}$	-0.06	-1.12	-0.93	1.08	0.59	0.10	-0.79
		$\hat{\rho}_A$	-0.02	-1.04	-1.61	1.74	0.55	0.19	-0.63
		$\hat{\rho}_{A(M)}$	-0.05	-1.09	-1.56	1.62	0.57	0.17	-0.65
		$\hat{\rho}_{\kappa}$	0.05	-0.90	-4.87	6.12	0.50	0.69	0.37
		$\hat{\rho}_{\kappa(M)}$	0.03	-0.94	-3.87	4.93	0.51	0.56	0.13
120	10	$\hat{\kappa}$	-0.04	-1.07	-0.50	0.81	0.60	0.06	-0.89
		$\hat{\rho}_A$	0.00	-1.00	-0.76	1.26	0.53	0.11	-0.78
		$\hat{\rho}_{A(M)}$	-0.02	-1.04	-0.74	1.18	0.57	0.10	-0.80
		$\hat{\rho}_{\kappa}$	0.04	-0.91	-5.06	9.51	0.50	0.57	0.14
		$\hat{\rho}_{\kappa(M)}$	0.03	-0.94	-3.73	6.99	0.52	0.44	-0.11
	20	$\hat{\kappa}$	-0.02	-1.05	-0.38	0.42	0.60	0.04	-0.91
		$\hat{\rho}_A$	-0.01	-1.01	-0.42	0.83	0.53	0.07	-0.85
		$\hat{\rho}_{A(M)}$	-0.02	-1.03	-0.45	0.81	0.55	0.07	-0.86
		$\hat{\rho}_{\kappa}$	-0.04	-1.08	-3.76	3.27	0.51	0.37	-0.25
		$\hat{\rho}_{\kappa(M)}$	-0.03	-1.06	-2.94	2.74	0.51	0.29	-0.41

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.8: Properties of ICC estimators: based on 1000 simulations of trials with n clusters of size μ per group, cumulative odds ratio $\theta = 1.2$, intracluster correlation $\rho = 0$, and variable cluster size $\lambda = 0.8$

Parameters			Descriptive statistics					Descriptive statistics (setting negative estimates to zero)	
μ	n	ρ [1]	Average (× 100)	Relative bias	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)	Relative bias
50	10	$\hat{\kappa}$	0.01	-0.98	-1.43	2.00	0.50	0.20	-0.61
		$\hat{\rho}_A$	-0.01	-1.02	-1.79	2.71	0.54	0.26	-0.48
		$\hat{\rho}_{A(M)}$	-0.06	-1.11	-2.02	2.99	0.56	0.24	-0.51
		$\hat{\rho}_{\kappa}$	-0.04	-1.08	-8.41	6.67	0.51	0.92	0.83
		$\hat{\rho}_{\kappa(M)}$	-0.05	-1.10	-6.78	6.09	0.52	0.76	0.52
	20	$\hat{\kappa}$	0.07	-0.86	-0.79	1.45	0.44	0.17	-0.66
		$\hat{\rho}_A$	-0.01	-1.02	-1.52	1.73	0.53	0.18	-0.63
		$\hat{\rho}_{A(M)}$	-0.03	-1.07	-1.52	1.56	0.56	0.17	-0.66
		$\hat{\rho}_{\kappa}$	-0.08	-1.16	-5.21	5.98	0.54	0.62	0.24
		$\hat{\rho}_{\kappa(M)}$	-0.08	-1.15	-4.18	5.46	0.53	0.52	0.03
120	10	$\hat{\kappa}$	0.08	-0.84	-0.55	1.25	0.38	0.13	-0.74
		$\hat{\rho}_A$	0.00	-0.99	-0.77	1.19	0.53	0.11	-0.77
		$\hat{\rho}_{A(M)}$	-0.02	-1.03	-0.76	1.28	0.56	0.11	-0.79
		$\hat{\rho}_{\kappa}$	0.00	-1.00	-4.55	5.37	0.50	0.55	0.10
		$\hat{\rho}_{\kappa(M)}$	0.00	-1.01	-3.63	4.27	0.50	0.45	-0.11
	20	$\hat{\kappa}$	0.12	-0.77	-0.28	0.66	0.24	0.14	-0.73
		$\hat{\rho}_A$	0.01	-0.99	-0.53	0.86	0.50	0.08	-0.83
		$\hat{\rho}_{A(M)}$	0.00	-1.01	-0.54	0.84	0.52	0.08	-0.84
		$\hat{\rho}_{\kappa}$	0.03	-0.94	-3.44	2.90	0.50	0.39	-0.21
		$\hat{\rho}_{\kappa(M)}$	0.02	-0.96	-2.91	2.53	0.49	0.32	-0.36

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.9: Properties of ICC estimators: based on 1000 simulations of trials with n clusters of size μ per group, cumulative odds ratio $\theta = 1$, intracluster correlation $\rho = 0.005$, and variable cluster size $\lambda = 0.8$

Parameters			Descriptive statistics					Descriptive statistics (setting negative estimates to zero)	
μ	n	ρ [1]	Average (× 100)	Relative bias	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)	Relative bias
50	10	$\hat{\kappa}$	0.36	-0.29	-1.48	3.29	0.28	0.43	-0.13
		$\hat{\rho}_A$	0.47	-0.06	-1.73	4.91	0.32	0.61	0.21
		$\hat{\rho}_{A(M)}$	0.42	-0.16	-1.62	4.87	0.34	0.58	0.15
		$\hat{\rho}_{\kappa}$	0.33	-0.34	-9.04	7.68	0.45	1.13	1.26
		$\hat{\rho}_{\kappa(M)}$	0.36	-0.28	-7.17	6.72	0.44	0.98	0.96
	20	$\hat{\kappa}$	0.42	-0.15	-0.58	1.96	0.16	0.45	-0.10
		$\hat{\rho}_A$	0.49	-0.03	-0.94	2.84	0.22	0.54	0.08
		$\hat{\rho}_{A(M)}$	0.46	-0.07	-0.87	2.72	0.22	0.52	0.04
		$\hat{\rho}_{\kappa}$	0.43	-0.14	-4.84	7.08	0.41	0.93	0.87
		$\hat{\rho}_{\kappa(M)}$	0.43	-0.13	-4.09	6.10	0.40	0.81	0.63
120	10	$\hat{\kappa}$	0.44	-0.12	-0.35	1.69	0.01	0.45	-0.11
		$\hat{\rho}_A$	0.51	0.01	-0.58	2.47	0.14	0.53	0.06
		$\hat{\rho}_{A(M)}$	0.49	-0.02	-0.59	2.49	0.16	0.52	0.04
		$\hat{\rho}_{\kappa}$	0.37	-0.25	-6.53	6.44	0.43	0.85	0.69
		$\hat{\rho}_{\kappa(M)}$	0.39	-0.21	-4.52	4.90	0.40	0.72	0.45
	20	$\hat{\kappa}$	0.46	-0.07	-0.10	1.34	0.01	0.46	-0.07
		$\hat{\rho}_A$	0.51	0.01	-0.25	1.55	0.04	0.51	0.02
		$\hat{\rho}_{A(M)}$	0.50	-0.01	-0.30	1.55	0.04	0.50	0.00
		$\hat{\rho}_{\kappa}$	0.45	-0.10	-2.93	4.18	0.34	0.70	0.41
		$\hat{\rho}_{\kappa(M)}$	0.46	-0.08	-2.11	3.42	0.29	0.62	0.25

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.10: Properties of ICC estimators: based on 1000 simulations of trials with n clusters of size μ per group, cumulative odds ratio $\theta = 1.2$, intracluster correlation $\rho = 0.005$, and variable cluster size $\lambda = 0.8$

Parameters			Descriptive statistics					Descriptive statistics (setting negative estimates to zero)	
μ	n	ρ [1]	Average (× 100)	Relative bias	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)	Relative bias
50	10	$\hat{\kappa}$	0.48	-0.03	-1.37	3.90	0.21	0.54	0.09
		$\hat{\rho}_A$	0.46	-0.09	-1.81	3.70	0.32	0.59	0.18
		$\hat{\rho}_{A(M)}$	0.42	-0.16	-1.90	3.86	0.33	0.57	0.14
		$\hat{\rho}_{\kappa}$	0.37	-0.27	-7.85	13.72	0.43	1.14	1.28
		$\hat{\rho}_{\kappa(M)}$	0.38	-0.23	-6.48	10.45	0.42	1.01	1.02
	20	$\hat{\kappa}$	0.56	0.12	-0.62	2.07	0.07	0.57	0.14
		$\hat{\rho}_A$	0.50	0.01	-1.03	3.14	0.20	0.55	0.11
		$\hat{\rho}_{A(M)}$	0.48	-0.04	-0.97	3.14	0.21	0.54	0.07
		$\hat{\rho}_{\kappa}$	0.46	-0.08	-5.62	7.01	0.40	0.95	0.90
		$\hat{\rho}_{\kappa(M)}$	0.47	-0.05	-4.82	6.11	0.38	0.86	0.73
120	10	$\hat{\kappa}$	0.55	0.10	-0.26	2.44	0.03	0.55	0.11
		$\hat{\rho}_A$	0.52	0.04	-0.68	2.44	0.12	0.54	0.07
		$\hat{\rho}_{A(M)}$	0.50	0.00	-0.68	2.63	0.13	0.52	0.04
		$\hat{\rho}_{\kappa}$	0.56	0.11	-4.22	6.32	0.36	0.91	0.82
		$\hat{\rho}_{\kappa(M)}$	0.54	0.08	-3.49	4.80	0.33	0.80	0.61
	20	$\hat{\kappa}$	0.58	0.15	-0.13	1.54	0.01	0.58	0.15
		$\hat{\rho}_A$	0.51	0.01	-0.32	1.79	0.05	0.51	0.02
		$\hat{\rho}_{A(M)}$	0.49	-0.01	-0.36	1.82	0.06	0.50	0.00
		$\hat{\rho}_{\kappa}$	0.49	-0.01	-4.24	4.31	0.33	0.74	0.48
		$\hat{\rho}_{\kappa(M)}$	0.49	-0.01	-3.40	3.83	0.29	0.67	0.34

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.11: Properties of ICC estimators: based on 1000 simulations of trials with n clusters of size μ per group, cumulative odds ratio $\theta = 1$, intracluster correlation $\rho = 0.01$, and variable cluster size $\lambda = 0.8$

Parameters			Descriptive statistics					Descriptive statistics (setting negative estimates to zero)	
μ	n	ρ [1]	Average (× 100)	Relative bias	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)	Relative bias
50	10	$\hat{\kappa}$	0.86	-0.14	-0.91	3.16	0.10	0.88	-0.12
		$\hat{\rho}_A$	1.02	0.02	-1.38	5.87	0.17	1.08	0.08
		$\hat{\rho}_{A(M)}$	0.97	-0.03	-1.42	6.30	0.19	1.05	0.05
		$\hat{\rho}_{\kappa}$	0.93	-0.07	-8.12	14.54	0.37	1.57	0.57
		$\hat{\rho}_{\kappa(M)}$	0.94	-0.06	-6.55	12.02	0.35	1.41	0.41
	20	$\hat{\kappa}$	0.90	-0.10	-0.35	2.69	0.02	0.91	-0.09
		$\hat{\rho}_A$	0.97	-0.03	-0.75	3.64	0.07	0.98	-0.02
		$\hat{\rho}_{A(M)}$	0.95	-0.05	-0.87	3.75	0.07	0.97	-0.03
		$\hat{\rho}_{\kappa}$	0.91	-0.09	-5.80	7.25	0.33	1.30	0.30
		$\hat{\rho}_{\kappa(M)}$	0.93	-0.07	-4.51	6.22	0.28	1.20	0.20
120	10	$\hat{\kappa}$	0.88	-0.12	-0.08	2.38	0.00	0.88	-0.12
		$\hat{\rho}_A$	0.98	-0.02	-0.46	3.26	0.03	0.99	-0.01
		$\hat{\rho}_{A(M)}$	0.96	-0.04	-0.50	3.42	0.04	0.96	-0.04
		$\hat{\rho}_{\kappa}$	0.81	-0.19	-4.08	8.13	0.32	1.18	0.18
		$\hat{\rho}_{\kappa(M)}$	0.84	-0.16	-3.02	5.65	0.27	1.07	0.07
	20	$\hat{\kappa}$	0.95	-0.05	0.17	2.06	0.00	0.95	-0.05
		$\hat{\rho}_A$	1.00	0.00	-0.06	2.75	0.00	1.00	0.00
		$\hat{\rho}_{A(M)}$	0.98	-0.02	-0.11	2.76	0.00	0.98	-0.02
		$\hat{\rho}_{\kappa}$	0.97	-0.03	-3.56	4.61	0.23	1.13	0.13
		$\hat{\rho}_{\kappa(M)}$	0.96	-0.04	-2.29	4.22	0.16	1.05	0.05

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.12: Properties of ICC estimators: based on 1000 simulations of trials with n clusters of size μ per group, cumulative odds ratio $\theta = 1.2$, intracluster correlation $\rho = 0.01$, and variable cluster size $\lambda = 0.8$

Parameters			Descriptive statistics					Descriptive statistics (setting negative estimates to zero)	
μ	n	ρ [1]	Average (× 100)	Relative bias	Minimum (× 100)	Maximum (× 100)	Percentage of negative values	Average (× 100)	Relative bias
50	10	$\hat{\kappa}$	0.96	-0.04	-1.31	4.06	0.09	0.98	-0.02
		$\hat{\rho}_A$	0.98	-0.02	-1.38	5.29	0.18	1.05	0.05
		$\hat{\rho}_{A(M)}$	0.93	-0.07	-1.42	5.14	0.19	1.01	0.01
		$\hat{\rho}_{\kappa}$	0.94	-0.06	-8.24	10.83	0.37	1.53	0.53
		$\hat{\rho}_{\kappa(M)}$	0.93	-0.07	-7.06	10.05	0.35	1.40	0.40
	20	$\hat{\kappa}$	1.07	0.07	-0.30	2.87	0.01	1.07	0.07
		$\hat{\rho}_A$	1.02	0.02	-0.83	3.86	0.06	1.04	0.04
		$\hat{\rho}_{A(M)}$	0.99	-0.01	-0.76	3.73	0.07	1.01	0.01
		$\hat{\rho}_{\kappa}$	0.97	-0.03	-7.76	8.38	0.31	1.33	0.33
		$\hat{\rho}_{\kappa(M)}$	0.95	-0.05	-6.88	7.28	0.29	1.24	0.24
120	10	$\hat{\kappa}$	1.01	0.01	-0.25	2.69	0.00	1.01	0.01
		$\hat{\rho}_A$	1.00	0.00	-0.88	4.18	0.05	1.00	0.00
		$\hat{\rho}_{A(M)}$	0.98	-0.02	-0.75	5.34	0.01	0.98	-0.02
		$\hat{\rho}_{\kappa}$	0.91	-0.09	-5.80	6.21	0.29	1.22	0.22
		$\hat{\rho}_{\kappa(M)}$	0.92	-0.08	-4.93	5.57	0.25	1.13	0.13
	20	$\hat{\kappa}$	1.06	0.06	0.21	2.23	0.00	1.06	0.06
		$\hat{\rho}_A$	0.97	-0.03	-0.09	2.61	0.00	0.97	-0.03
		$\hat{\rho}_{A(M)}$	0.96	-0.04	-0.10	2.71	0.00	0.96	-0.04
		$\hat{\rho}_{\kappa}$	0.91	-0.09	-2.56	4.56	0.24	1.05	0.05
		$\hat{\rho}_{\kappa(M)}$	0.91	-0.09	-2.07	4.06	0.19	0.99	-0.01

[1] ICC estimators $\hat{\kappa}$, $\hat{\rho}_A$, $\hat{\rho}_{A(M)}$, $\hat{\rho}_{\kappa}$, and $\hat{\rho}_{\kappa(M)}$ were denoted in Section 6.4.1 and Table 6.3.

Table 7.13: Type I error rates of adjusted Cochran-Armitage test statistics1: based on 1000 simulations of trials with *n* clusters of size μ per group, cumulative odds ratio θ, intracluster correlation ρ, and fixed cluster sizes (overly liberal or conservative type I error rates are in bold font)

Parameters			Adjusted test statistics[1]					
μ	n	ρ	χ^2_{A1}	$\chi^2_{\kappa 1}$	χ^2_{J}	$\chi^2_{A1(M)}$	$\chi^2_{\kappa 1(M)}$	$\chi^2_{J(M)}$
50	10	0	0.046	0.047	0.061	0.048	0.048	0.056
		0.005	0.050	0.051	0.051	0.047	0.048	0.051
		0.01	0.056	0.058	0.047	0.050	0.053	0.050
	20	0	0.033	0.034	0.040	0.031	0.032	0.036
		0.005	0.051	0.052	0.052	0.051	0.054	0.049
		0.01	0.047	0.047	0.044	0.044	0.046	0.041
120	10	0	0.042	0.044	0.047	0.038	0.038	0.048
		0.005	0.059	0.065	0.051	0.053	0.056	0.050
		0.01	0.056	0.058	0.045	0.052	0.058	0.044
	20	0	0.044	0.044	0.056	0.049	0.049	0.061
		0.005	0.051	0.051	0.046	0.052	0.053	0.045
		0.01	0.054	0.057	0.051	0.056	0.058	0.050

[1] Adjusted test statistics χ^2_{A1}, $\chi^2_{\kappa 1}$, χ^2_{J}, $\chi^2_{A1(M)}$, $\chi^2_{\kappa 1(M)}$, and $\chi^2_{J(M)}$ were denoted in Section 6.4.2 and Table 6.4, and negative ICC estimators in the calculation of adjusted test statistics were set to zero

Table 7.14: Type I error rates of adjusted Cochran-Armitage test statistics1: based on 1000 simulations of trials with *n* clusters of size μ per group, intracluster correlation ρ, and variable cluster size $\lambda = 0.8$ (overly liberal or conservative type I error rates are in bold font)

Parameters			Adjusted test statistics[1]													
μ	n	ρ	χ^2_{A1}	χ^2_{A2}	χ^2_{A3}	$\chi^2_{\kappa 1}$	$\chi^2_{\kappa 2}$	$\chi^2_{\kappa 3}$	$\chi^2_{A1(M)}$	$\chi^2_{A2(M)}$	$\chi^2_{A3(M)}$	$\chi^2_{\kappa 1(M)}$	$\chi^2_{\kappa 2(M)}$	$\chi^2_{\kappa 3(M)}$	χ^2_{J}	$\chi^2_{J(M)}$
50	10	0	0.046	0.046	0.046	0.041	0.041	0.041	0.046	0.046	0.046	0.036	0.036	0.036	0.059	0.058
		0.005	0.05	0.05	0.051	0.057	0.057	0.058	0.056	0.056	0.056	0.06	0.06	0.06	0.058	0.058
		0.01	**0.074**	**0.074**	**0.077**	**0.077**	**0.077**	**0.078**	**0.071**	**0.071**	**0.073**	**0.073**	**0.073**	**0.076**	0.06	0.058
	20	0	0.044	0.044	0.044	0.034	0.034	0.034	0.042	0.042	0.044	0.032	0.032	0.032	0.054	0.052
		0.005	0.06	0.06	0.06	0.051	0.051	0.053	0.071	0.071	0.072	0.062	0.062	0.064	0.061	0.065
		0.01	0.06	0.06	0.062	**0.071**	**0.071**	**0.071**	0.052	0.052	0.055	0.057	0.057	0.059	0.057	0.049
120	10	0	0.04	0.04	0.042	**0.025**	**0.025**	**0.025**	0.043	0.043	0.045	0.032	0.032	0.032	0.053	0.056
		0.005	0.057	0.057	0.057	0.064	0.064	0.064	0.056	0.056	0.058	0.063	0.063	0.063	0.048	0.045
		0.01	0.062	0.062	0.067	**0.104**	**0.104**	**0.106**	0.058	0.059	0.064	**0.094**	**0.094**	**0.095**	0.047	0.044
	20	0	0.042	0.042	0.042	**0.03**	**0.03**	**0.03**	0.043	0.043	0.045	**0.031**	**0.031**	0.032	0.05	0.053
		0.005	0.054	0.054	0.06	**0.069**	**0.069**	**0.069**	0.061	0.061	0.063	**0.07**	**0.07**	**0.071**	0.055	0.052
		0.01	0.048	0.048	0.053	**0.082**	**0.082**	**0.083**	0.051	0.051	0.054	**0.077**	**0.077**	**0.078**	0.04	0.041

[1] Adjusted test statistics were denoted in Section 6.4.2 and Table 6.4, and negative ICC estimators in the calculation of adjusted test statistics were set to zero.

Table 7.15: Power of adjusted Cochran-Armitage test statistics: based on 1000 simulations of trials with *n* clusters of size *μ* per group, intracluster correlation *ρ*, and fixed cluster size *λ* = 1

Parameters			Adjusted test statistics[1]					
μ	n	ρ	χ^2_{A1}	χ^2_{K1}	χ^2_J	χ'^2_{A1}	χ'^2_{K1}	χ'^2_J
50	10	0	0.305	0.307	0.325	0.300	0.306	0.326
		0.005	0.293	0.299	0.282	0.302	0.308	0.287
		0.01	0.245	0.260	0.215	0.257	0.263	0.221
	20	0	0.557	0.558	0.577	0.554	0.556	0.577
		0.005	0.506	0.510	0.496	0.507	0.511	0.498
		0.01	0.419	0.423	0.398	0.425	0.432	0.409
120	10	0	0.608	0.610	0.602	0.618	0.620	0.606
		0.005	0.492	0.503	0.456	0.500	0.506	0.463
		0.01	0.363	0.382	0.337	0.375	0.393	0.333
	20	0	0.918	0.918	0.922	0.921	0.922	0.928
		0.005	0.745	0.751	0.728	0.760	0.765	0.742
		0.01	0.600	0.611	0.584	0.610	0.625	0.587

[1] Adjusted test statistics were denoted in Section 6.4.2 and Table 6.5, and negative ICC estimators in the calculation of adjusted test statistics were set to zero.

Table 7.16: Power of adjusted Cochran-Armitage test statistics: based on 1000 simulations of trials with n clusters of size μ per group, intracluster correlation ρ, and variable cluster size $\lambda = 0.8$

Parameters			Adjusted test statistics[1]													
μ	n	ρ	χ^2_{A1}	χ^2_{A2}	χ^2_{A3}	$\chi^2_{\kappa 1}$	$\chi^2_{\kappa 2}$	$\chi^2_{\kappa 3}$	χ'^2_{A1}	χ'^2_{A2}	χ'^2_{A3}	$\chi'^2_{\kappa 1}$	$\chi'^2_{\kappa 2}$	$\chi'^2_{\kappa 3}$	χ'^2_{J}	χ'^2_{J}
50	10	0	0.292	0.292	0.292	0.236	0.236	0.238	0.3	0.3	0.3	0.252	0.252	0.255	0.317	0.312
		0.005	0.281	0.281	0.283	0.234	0.234	0.238	0.283	0.283	0.285	0.236	0.236	0.238	0.254	0.27
		0.01	0.256	0.256	0.26	0.232	0.231	0.238	0.253	0.253	0.255	0.239	0.239	0.247	0.232	0.231
	20	0	0.562	0.562	0.563	0.481	0.481	0.485	0.578	0.578	0.578	0.496	0.496	0.506	0.593	0.599
		0.005	0.506	0.506	0.509	0.442	0.442	0.448	0.501	0.501	0.503	0.451	0.451	0.457	0.499	0.494
		0.01	0.412	0.411	0.423	0.405	0.405	0.408	0.413	0.413	0.423	0.413	0.413	0.42	0.386	0.384
120	10	0	0.598	0.598	0.599	0.468	0.468	0.475	0.605	0.605	0.606	0.502	0.502	0.505	0.609	0.611
		0.005	0.447	0.445	0.454	0.396	0.396	0.406	0.453	0.453	0.465	0.419	0.418	0.428	0.391	0.395
		0.01	0.345	0.343	0.355	0.342	0.342	0.36	0.338	0.338	0.352	0.356	0.356	0.369	0.307	0.299
	20	0	0.891	0.891	0.891	0.799	0.799	0.81	0.894	0.894	0.894	0.826	0.826	0.83	0.889	0.894
		0.005	0.727	0.727	0.738	0.668	0.668	0.684	0.732	0.732	0.741	0.685	0.684	0.695	0.71	0.713
		0.01	0.561	0.561	0.579	0.568	0.567	0.584	0.571	0.571	0.585	0.58	0.58	0.606	0.549	0.551

[1] Adjusted test statistics were denoted in Section 6.4.2 and Table 6.5, and negative ICC estimators in the calculation of adjusted test statistics were set to zero.

Table 7.17: Type I error rates of model-based test statistics: based on 1000 simulations of trials with *n* clusters of size *μ* per group, intracluster correlation *ρ*, and fixed cluster size *λ* = 0.8. (overly liberal or conservative type I error rates are in bold font)

Parameters			Model-based test statistics[1]														
μ	n	ρ	W_M	W_R	W_{BC1}	W_{BC2}	W_{BC3}	W_{BC4}	W_{BC5}	W_{df1}	W_{df2}	W_{df3}	W_{df4}	W_{df5}	S_R	S_{BC}	χ^2_{CS}	T_{Linear}
50	10	0	0.048	**0.077**	0.052	0.061	0.056	0.052	0.055	0.054	0.056	0.057	0.057	0.047	0.053	0.055	0.031	0.039
		0.005	**0.072**	0.064	0.047	0.054	0.050	0.047	0.052	0.050	0.050	0.050	0.050	0.037	0.048	0.051	0.039	0.050
		0.01	**0.105**	**0.084**	0.057	0.067	0.063	0.057	0.063	0.061	0.063	0.063	0.063	0.052	0.057	0.063	0.054	0.059
	20	0	0.045	0.057	0.048	0.053	0.053	0.048	0.052	0.051	0.053	0.052	0.052	0.046	0.049	0.052	0.035	0.034
		0.005	**0.082**	**0.070**	0.059	0.059	0.059	0.059	0.058	0.059	0.059	0.059	0.059	0.054	0.059	0.059	0.050	0.053
		0.01	**0.111**	**0.072**	0.055	0.062	0.062	0.055	0.068	0.061	0.062	0.062	0.062	0.051	0.056	0.060	0.060	0.067
120	10	0	0.055	**0.077**	0.040	0.059	0.055	0.040	0.038	0.047	0.050	0.052	0.052	0.029	0.040	0.051	0.047	0.041
		0.005	**0.103**	0.062	0.041	0.056	0.051	0.041	0.040	0.047	0.048	0.050	0.050	0.028	0.040	0.048	0.051	0.044
		0.01	**0.181**	**0.082**	0.061	0.073	0.065	0.059	0.058	0.064	0.066	0.068	0.068	0.050	0.059	0.067	0.069	0.068
	20	0	0.043	0.063	0.056	0.060	0.058	0.056	0.056	0.055	0.055	0.055	0.055	0.047	0.053	0.055	0.050	0.041
		0.005	**0.109**	0.063	0.050	0.056	0.054	0.050	0.046	0.054	0.054	0.054	0.054	0.044	0.051	0.054	0.054	0.055
		0.01	**0.190**	**0.074**	0.054	0.062	0.062	0.054	0.049	0.056	0.060	0.060	0.060	0.049	0.055	0.061	0.065	0.065

[1] Model-based test statistics were denoted in Section 6.4.2 and Table 6.5.

Table 7.18: Type I error rates of model-based test statistics: based on 1000 simulations of trials with *n* clusters of size *μ* per group, intracluster correlation *ρ*, and variable cluster size *λ* = 0.8 (overly liberal or conservative type I error rates are in bold font)

Parameters			Model-based test statistics[1]														
μ	n	ρ	W_M	W_R	W_{BC1}	W_{BC2}	W_{BC3}	W_{BC4}	W_{BC5}	W_{df1}	W_{df2}	W_{df3}	W_{df4}	W_{df5}	S_R	S_{BC}	χ^2_{CS}	T_{Linear}
50	10	0	0.052	0.084	0.049	0.061	0.056	0.049	0.052	0.056	0.058	**0.016**	**0.016**	0.045	0.043	0.050	**0.030**	0.038
		0.005	**0.082**	**0.099**	0.058	**0.075**	0.066	0.058	0.068	0.066	**0.071**	**0.016**	**0.016**	0.055	0.056	0.066	0.055	0.059
		0.01	**0.121**	**0.101**	0.065	**0.081**	**0.076**	0.065	**0.072**	**0.078**	**0.079**	**0.017**	**0.017**	0.055	0.055	0.068	0.065	**0.073**
	20	0	0.059	**0.079**	0.058	0.068	0.064	0.058	0.068	0.065	0.067	0.038	0.038	0.054	0.056	0.059	0.044	0.050
		0.005	**0.084**	0.065	0.052	0.060	0.058	0.052	0.055	0.057	0.057	0.032	0.032	0.047	0.047	0.049	0.051	0.059
		0.01	**0.114**	0.057	0.049	0.050	0.050	0.049	0.055	0.049	0.049	**0.029**	**0.029**	0.047	0.046	0.049	0.054	0.053
120	10	0	0.050	**0.101**	0.063	**0.080**	**0.077**	0.063	0.056	**0.073**	**0.077**	**0.018**	**0.018**	0.055	0.057	0.065	0.036	0.040
		0.005	**0.139**	**0.093**	0.058	**0.080**	**0.074**	0.058	0.059	**0.072**	**0.073**	**0.016**	**0.016**	0.053	0.054	0.062	0.052	0.057
		0.01	**0.216**	**0.101**	0.067	**0.088**	**0.082**	0.067	0.062	**0.083**	**0.084**	**0.027**	**0.027**	0.062	0.065	**0.069**	**0077**	**0.081**
	20	0	0.041	0.062	0.056	0.062	0.060	0.056	0.033	0.052	0.052	**0.028**	**0.028**	0.045	0.047	0.049	0.033	0.038
		0.005	**0.147**	**0.075**	0.068	**0.080**	**0.074**	0.068	0.055	0.067	0.068	0.036	0.036	0.054	0.058	0.064	0.063	0.068
		0.01	**0.231**	0.062	0.051	0.057	0.055	0.051	0.051	0.047	0.050	**0.029**	**0.029**	0.046	0.045	0.046	0.054	0.060

[1]Model-based test statistics were denoted in Section 6.4.2 and Table 6.5.

Table 7.19: Power of model-based test statistics: based on 1000 simulations of trials with *n* clusters of size μ per group, intracluster correlation ρ, and fixed cluster size $\lambda = 1$

Parameters			Model-based test statistics[1]												
μ	n	ρ	W_{BC1}	W_{BC2}	W_{BC3}	W_{BC4}	W_{BC5}	W_{df1}	W_{df2}	W_{df3}	W_{df4}	W_{df5}	S_R	S_{BC}	χ^2_{CS}	χ^2_l
50	10	0	0.304	0.347	0.337	0.303	0.263	0.324	0.333	0.332	0.332	0.267	0.305	0.331	0.254	0.277
		0.005	0.262	0.295	0.284	0.262	0.262	0.276	0.279	0.281	0.281	0.242	0.263	0.280	0.259	0.270
		0.01	0.250	0.288	0.274	0.250	0.253	0.263	0.268	0.270	0.270	0.225	0.256	0.271	0.262	0.275
	20	0	0.579	0.598	0.591	0.579	0.571	0.590	0.593	0.593	0.593	0.570	0.580	0.591	0.550	0.554
		0.005	0.518	0.541	0.535	0.518	0.510	0.527	0.533	0.534	0.534	0.503	0.520	0.531	0.525	0.532
		0.01	0.405	0.435	0.426	0.405	0.408	0.415	0.424	0.427	0.427	0.395	0.410	0.421	0.426	0.430
120	10	0	0.621	0.657	0.646	0.621	0.590	0.636	0.643	0.643	0.643	0.587	0.621	0.644	0.561	0.612
		0.005	0.447	0.484	0.472	0.447	0.439	0.461	0.567	0.467	0.467	0.400	0.446	0.466	0.455	0.483
		0.01	0.350	0.376	0.368	0.349	0.351	0.362	0.365	0.364	0.364	0.317	0.349	0.365	0.368	0.373
	20	0	0.908	0.915	0.913	0.908	0.906	0.911	0.913	0.913	0.913	0.901	0.908	0.912	0.898	0.910
		0.005	0.736	0.748	0.747	0.735	0.736	0.744	0.747	0.746	0.746	0.718	0.736	0.745	0.753	0.752
		0.01	0.602	0.617	0.611	0.602	0.600	0.608	0.611	0.610	0.610	0.591	0.602	0.610	0.618	0.617

[1]Model-based test statistics were denoted in Section 6.4.2 and Table 6.6.

Table 7.20: Power of model-based test statistics: based on 1000 simulations of trials with *n* clusters of size μ per group, intracluster correlation ρ, and variable cluster size $\lambda = 0.8$

Parameters			Model-based test statistics[1]							
μ	n	ρ	W_{BC1}	W_{BC4}	W_{BC5}	W_{df5}	S_R	S_{BC}	χ^2_{CS}	χ^2_l
50	10	0	0.309	0.309	0.286	0.293	0.300	0.328	0.266	0.295
		0.005	0.256	0.256	0.262	0.240	0.253	0.279	0.259	0.274
		0.01	0.230	0.230	0.242	0.222	0.227	0.250	0.241	0.250
	20	0	0.568	0.568	0.550	0.562	0.568	0.583	0.534	0.545
		0.005	0.455	0.455	0.454	0.444	0.451	0.463	0.457	0.465
		0.01	0.397	0.397	0.401	0.389	0.396	0.407	0.406	0.417
120	10	0	0.592	0.592	0.491	0.569	0.592	0.626	0.556	0.585
		0.005	0.397	0.397	0.371	0.378	0.384	0.409	0.439	0.446
		0.01	0.316	0.316	0.292	0.301	0.312	0.335	0.347	0.360
	20	0	0.903	0.903	0.814	0.899	0.902	0.908	0.888	0.896
		0.005	.0706	.0706	0.664	0.694	0.709	0.723	0.734	0.736
		0.01	0.562	0.562	0.545	0.552	0.561	0.570	0.601	0.594

[1]Model-based test statistics were denoted in Section 6.4.2 and Table 6.6.

Table 7.21: Regression Coefficient Estimates and their Standard Errors from marginal and cluster models: based on 1000 simulations of trials with n clusters of size μ per group, intracluster correlation ρ, and fixed cluster size $\lambda = 1$

	Parameters			Regression Coefficient Estimates		Standard Errors	
	μ	n	ρ	$\hat{\beta}_{GEE}$	$\hat{\beta}_{CS}$	$SE(\hat{\beta}_{GEE})$	$SE(\hat{\beta}_{CS})$
$\theta = 1$	50	10	0	0.001	0.000	0.112	0.124
			0.005	0.004	0.005	0.125	0.131
			0.01	0.003	0.002	0.136	0.141
		20	0	-0.005	-0.005	0.082	0.087
			0.005	-0.004	-0.004	0.091	0.093
			0.01	0.002	0.002	0.100	0.101
	120	10	0	-0.006	0.145	0.073	0.081
			0.005	-0.002	0.085	0.092	0.094
			0.01	-0.003	0.031	0.107	0.109
		20	0	0.000	0.134	0.053	0.058
			0.005	0.000	0.049	0.066	0.067
			0.01	-0.005	0.022	0.078	0.079
$\theta = 1.2$	50	10	0	0.182	0.182	0.112	0.124
			0.005	0.180	0.181	0.124	0.130
			0.01	0.190	0.191	0.136	0.140
		20	0	0.185	0.185	0.082	0.087
			0.005	0.194	0.194	0.091	0.093
			0.01	0.182	0.183	0.099	0.100
	120	10	0	0.183	0.236	0.072	0.079
			0.005	0.182	0.212	0.092	0.093
			0.01	0.184	0.201	0.107	0.108
		20	0	0.185	0.225	0.052	0.056
			0.005	0.180	0.193	0.066	0.066
			0.01	0.183	0.182	0.078	0.078

Table 7.22: Regression Coefficient Estimates and their Standard Errors from marginal and cluster models: based on 1000 simulations of trials with *n* clusters of size μ per group, intracluster correlation ρ, and variable cluster size $\lambda = 0.8$

	Parameters			Regression Coefficient Estimates		Standard Errors	
	μ	n	ρ	$\hat{\beta}_{GEE}$	$\hat{\beta}_{CS}$	$SE(\hat{\beta}_{GEE})$	$SE(\hat{\beta}_{CS})$
$\theta = 1$	50	10	0	-0.005	-0.005	0.111	0.126
			0.005	0.000	0.001	0.126	0.134
			0.01	0.002	0.002	0.139	0.144
		20	0	-0.002	-0.002	0.082	0.088
			0.005	-0.002	-0.002	0.093	0.095
			0.01	-0.002	-0.003	0.104	0.105
	120	10	0	0.000	0.000	0.072	0.081
			0.005	0.003	0.003	0.093	0.096
			0.01	-0.008	-0.007	0.110	0.111
		20	0	-0.001	-0.001	0.053	0.057
			0.005	-0.001	-0.001	0.069	0.069
			0.01	-0.002	-0.003	0.082	0.082
$\theta = 1.2$	50	10	0	0.182	0.182	0.111	0.126
			0.005	0.184	0.185	0.126	0.134
			0.01	0.188	0.190	0.139	0.144
		20	0	0.183	0.183	0.080	0.087
			0.005	0.182	0.182	0.092	0.094
			0.01	0.185	0.186	0.102	0.103
	120	10	0	0.180	0.280	0.071	0.081
			0.005	0.182	0.183	0.093	0.096
			0.01	0.184	0.185	0.110	0.111
		20	0	0.184	0.184	0.053	0.056
			0.005	0.183	0.193	0.068	0.068
			0.01	0.183	0.185	0.082	0.081

Chapter 8

8 Example: A school-based smoking prevention cluster randomization trial

8.1 Introduction

In this chapter, we use data from a school-based smoking prevention study to illustrate application of methods described in previous chapters. The Television School and Family Smoking Prevention and Cessation Project (TVSFP) is a cluster randomization trial, which was designed to test the independent and combined effects of a classroom curriculum and television programming for social resistance skills training, smoking prevention, and smoking cessation (Flay, et al., 1988).

The initial study was conducted from 1986 to 1988. It consisted of 7351 students in seventh grade in 340 classrooms within 47 schools from Los Angeles and San Diego. Students were randomized to five study conditions: 1) a social-resistance (SR) classroom curriculum, 2) a TV intervention, 3) a health-information-base attention-control curriculum, 4) a SR curriculum combined with a TV intervention (SR+TV), and 5) a no-intervention group. Randomization for this study was at the school level while the intervention was delivered to students in the classroom.

For this illustration, a subset of the TVSFP data was used. This subset included 1600 students from 135 classrooms and 28 Los Angeles schools. A tobacco and health knowledge scale (THKS) score was one of the primary study outcome variables and the one chosen for this study. The score was defined as the number of correct answers to seven questions on tobacco and health knowledge. According to Hedeker and Gibbons's (1994) study, three ordinal classifications were created for illustrative purposes, corresponding to 0-1, 2-3, and 4-7 correct answers. We further categorize the original study conditions into two groups: a TV intervention group (TV=yes) vs. non- TV group (TV=no). Moreover, our analysis will be limited to inferences about the effect of this school-based intervention on the ordinal THKS score.

The same data set was previously analyzed using a mixed effects model for ordinal outcomes (Hedeker and Gibbons, 1996) as well as binary outcomes (Gibbons and Hedeker, 1997). Also, Hedeker and Gibbons (1994) and Raman and Hedeker (2005) fit the data with mixed-effects ordinal probit and logistic regressions respectively. However, their investigations only focused on the analysis of cluster-specific models. In addition, their studies investigated the effects of all four conditions (SR, TV, TV+SR, and the non-intervention group) with the outcome THKS scores classified into four ordinal categories corresponding to 0-1, 2, 3, and 4-7 correct answers.

The rest of the chapter is organized as follows. Section 8.2 reviews the methods applied to the example data and Section 8.3 describes the results of the analysis. Conclusions and discussion are presented in Section 8.4.

8.2 Methods

Several summary statistics were calculated for the data. Table 8.1 shows descriptive statistics of the school size broken down by condition groups. In addition, student frequencies for three ordinal categories of the THKS are given in Table 8.2. The degree of imbalance in cluster size in each group was obtained as discussed in Chapter 6.

Estimates of the ICC were calculated for the THKS score among students within schools. These estimates were obtained by adapting one-way ANOVA and kappa-type methods as described in Sections 2.2 and 2.3.

Results from three adjusted Cochran-Armitage tests of the effect of TV intervention were compared and displayed in Table 8.4.Two types of ICC estimators were used to calculate the degree of variance inflation induced by clustering. In addition, results from thirteen model-based tests are compared and displayed in Table 8.5, including five bias-corrected and four degrees-of-freedom-adjusted approaches for the GEE Wald test and one corrected approach for the GEE score test. Comparisons between different methods of analysis focus on the statistical significance of associations between the TV intervention effect and the outcome THKS scores.

In addition, marginal and cluster-specific extensions of proportional odds models were compared in terms of strength of effect as measured by the magnitude of model parameter estimates and their standard errors. In particular, the marginal model was fitted by the GEE approach using an independent working correlation. The cluster-specific model was fitted by the Gauss-Quadrature approach. The SAS procedures PROC GENMOND and PROC NLMIXED (SAS V.9.2, SAS Institute, Inc, Cary, NC) were employed to fit the marginal and cluster-specific models respectively.

8.3 Results

8.3.1 Descriptive Analyses

In the TVSFP study, fourteen schools were randomized to each group. The descriptive statistics for the cluster (school) sizes in each group are listed in Table 8.1. School sizes in the non-TV group are more variable than those in the TV group.

The student frequencies for the three-category THKS scores are displayed in Table 8.2 for each group. The estimated cumulative odds ratio of the THKS scores comparing the TV group with the non-TV group is 0.966, which is close to one.

Table 8.1: Descriptive statistics of school size per intervention group in the TVSFP

Intervention group	Number of schools	Mean	Standard deviation	Minimum	Maximum	Imbalance
Non-TV group	21	57.2	38.7	23	137	0.69
TV group	7	57.1	22.1	18	94	0.87

Table 8.2: Frequencies of three-category THKS scores per intervention group (%)

Intervention Group	THKS score			
	0-1	2-3	4-7	Total
Non-TV group	179 (22.4)	402 (50.2)	220 (27.4)	801 (100%)
TV group	176 (22.0)	396 (49.6)	227 (28.4)	799 (100%)

8.3.2 ICC Estimation

Estimates of ICC for the THKS scores among students within schools are listed in Table 8.3. The two ANOVA ICC estimators are smaller than the two kappa-type ICC estimators. In addition, ICC estimators obtained by using scores 1, 2 and 3 are larger than those using midranks for both ANOVA and kappa-type estimators. This is probably due, in part, to the large discrepancies between the scores 1, 2, or 3 with midrank scores in this example. In particular, the midrank scores are 1, 4 and 8 for the three ordinal categories, as compared to scores 1, 2, and 3.

Table 8.3: Estimated ICCs for the THKS scores among students within schools

	ANOVA method using scores 1,2 or 3 ($\hat{\rho}_A$)	ANOVA method using midrank scores ($\hat{\rho}_{A(M)}$)	Kappa approach using scores 1,2 or 3 ($\hat{\rho}_\kappa$)	Kappa approach using midrank scores ($\hat{\rho}_{\kappa(M)}$)
ICC estimates	0.059	0.058	0.127	0.080

8.3.3 Adjusted Cochran-Armitage Tests

Three adjusted Cochran-Armitage tests using both ANOVA and kappa-type ICC estimators with midrank scores were applied to examine the effect of the TV intervention group. The corresponding six test statistics and their p-values are listed in Table 8.4. All test statistics and their p-values are quite similar to each other and indicate a non-significant TV program effect on the THKS scores. This generally agrees with the results reported earlier (e.g., Hedeker and Gibbons, 1994; Raman and Hedeker, 2005).

Table 8.4: Adjusted Cochran-Armitage test statistics for the effect of the TV intervention using ANOVA and kappa-type ICC estimators

Methods	Methods	Test statistics	Test statistics value (p-value)
Using ANOVA ICC estimator with midrank scores ($\hat{\rho}_{A(M)}$)	Donner and Donald's test	$\chi^2_{CO-(1)}$	0.026 (0.87)
	An Alternative to Donner and Donald's Test	$\chi^2_{CO-(2)}$	0.025(0.87)
	Weighted-Least-Square Cochran-Armitage Test	χ^2_{CO-WLS}	0.030(0.86)
Using kappa-type ICC estimator with midrank scores ($\hat{\rho}_{\kappa(M)}$)	Donner and Donald's test	$\chi^2_{CO-(1)}$	0.020(0.88)
	An Alternative to Donner and Donald's Test	$\chi^2_{CO-(2)}$	0.020(0.88)
	Weighted-Least-Square Cochran-Armitage Test	χ^2_{CO-WLS}	0.023(0.88)

8.3.4 Adjusted Model-based Tests

A marginal proportional odds model is now fit to the example data, where the THKS score is modeled in terms of a dummy-coded (no=0 and yes=1) TV effect. Thirteen model-based test statistics and their corresponding p-values which evaluate the TV invention effect are listed in Table 8.5.

All test results show a non-significant TV effect on the outcome THKS score. This conclusion is in agreement with previous reports (e.g., Hedeker and Gibbons, 1994; Raman and Hedeker, 2005). Adjusting and modifying robust tests did not affect inferences concerning the effect of the TV intervention program. However, the five sandwich bias correction approaches enlarged the robust variance estimates as we illustrated in Chapter 5. Also, the four degree-of-freedom-adjusted approaches slightly reduced the magnitude of inflated type I errors by adjusting the approximate F-test.

The test statistic obtained from the robust score test (S_R=0.0300 with p=0.8610) is smaller than that generated from the robust Wald test (W_R=0.0309 with p=8606). After adjusting, the modified score test has a slightly smaller p-value (p=0.8584) which is consistent with discussions in previous chapters.

Table 8.5: Test statistics for the TV intervention effect from the marginal extensions of cumulative logit model for the THKS scores (SAS procedure: PROC GENMOD)

Tests	Test statistic	Test statistics value (p-value)
Model-based Wald test	W_M	0.1333 (0.7153)
Robust Wald test	W_R	0.0309 (0.8606)
Robust score test	S_R	0.0300 (0.8610)
Bias-corrected Wald test: Approach 1	W_{BC1}	0.0250 (0.8745)
Bias-corrected Wald test: Approach 2	W_{BC2}	0.0277 (0.8677)
Bias-corrected Wald test: Approach 3	W_{BC3}	0.0271 (0.8693)
Bias-corrected Wald test: Approach 4	W_{BC4}	0.0250 (0.8745)
Bias-corrected Wald test: Approach 5	W_{BC5}	0.0267 (0.8702)
Degrees-of-freedom-adjusted Wald test: Approach 1	W_{df1}	0.0276 (0.8684)
Degrees-of-freedom-adjusted Wald test: Approach 2	W_{df2}	0.0309 (0.8622)
Degrees-of-freedom-adjusted Wald test: Approach 3	W_{df3}	0.0276 (0.8639)
Degrees-of-freedom-adjusted Wald test: Approach 4	W_{df4}	0.0309 (0.8606)
Degrees-of-freedom-adjusted Wald test: Approach 5	W_{df5}	0.0308 (0.8697)
Modified robust score test	S_{BC}	0.0300 (0.8584)

8.3.5 Relationship between marginal and cluster-specific models

Marginal and cluster specific models are now fitted to the example data. The parameter estimates and standard errors are listed in Table 8.6. The TV effect estimate is obtained as 0.0344 in the marginal model, which is close to the TV effect estimate obtained as 0.0166 in the cluster-specific model. In addition, the standard error of the TV effect estimate from the marginal model is 0.1958, which is smaller than the standard error 0.2096 from the cluster-specific model. This parallels results previously reported for binary data.

Table 8.6: Parameter estimates (log odds ratios) of the TV effect from marginal and mixed effects logistic regression models with cumulative logit for the THKS scores

Term	Log odds ratio in the marginal model (standard error)	Log odds ratio in the mixed effects model (standard error)
TV intervention	0.0344 (0.1958)	0.0166 (0.2096)

8.4 Discussion

Although mixed effects categorical modeling methods have been previously applied to the same example data, there are some differences. For example, previous studies (e.g.,Hedeker and Gibbons, 1994) considered that the schools were randomized to four study conditions (i.e., SR, TV, TV+SR, and the non-intervention group), while our study considered only two groups, i.e., the TV and non-TV group. In addition, previous studies (e.g., Raman and Hedeker, 2005) divided THKS scores into four ordinal classifications corresponding to 0-1, 2, 3, and 4-7 correct responses, while we grouped the THKS scores into three ordinal classifications. Some studies also considered cluster effects at both the class level and school level, while we considered cluster effects at the school level only. These differences may lead to some discrepancy in results between our study and previous studies. Also, previous studies (e.g., Raman and Hedeker, 2005) evaluated the intervention program effects while controlling for the baseline information. However, this research focuses on analysis with a single cluster-level covariate only, i.e., the TV intervention effect. Therefore, we did not consider the baseline smoking information here.

The ICC estimator within schools was 0.013 in Raman and Hedeker (2005)'s study (four category ordinal outcomes and three-level cluster effect at school and class level), 0.022 in Hedeker and Gibbon (1996)'s study (continuous outcomes and two-level cluster effect at clsss effect), and 0.026 in Gibbons and Hedeker (1997)'s study (three levels, binary). They are all smaller than the calculated values of the ANOVA and kappa ICC estimators. The reasons may be due to the different model variables in the current study and previous studies.

As discussed in Chapter 5, the GEE score test tends to have a smaller test size than nominal, in contrast to the liberal behaviour of the GEE Wald test. However, the score test statistic (S_R=0.0300) is slightly smaller the Wald test statistic (W_R=0.0309). This may be explained by research showing that the conservative behavior of robust score tests is reduced as the number of clusters increases (i.e., n=30) (Guo et al., 2005).

The regression coefficient estimates (log odds ratio) from both marginal and cluster-specific models are close to zero. Combined with the small ICC, this may explain why the two coefficient estimates are very similar to each other. The same reason may also explain why the analytic relationship discussed in Chapter 5 does not hold in this example.

Chapter 9

9 Conclusions

The primary objective of this thesis was to develop and evaluate methods that analyze correlated ordinal data obtained from cluster randomization trials. Attention was restricted to completely randomized community intervention trials assuming a single binary, cluster-level covariate. The purpose of this chapter is to summarize the most important findings of this thesis in Section 9.1, discuss potential limitations and suggest areas for future research in Section 9.2.

9.1 Summaries

9.1.1 Main Findings

Properties of methods compared used three approaches: algebraic computation, simulation and a case study. The complexity of most methods restricts algebraic comparisons to fairly simple situations where there are equal numbers of clusters with fixed cluster sizes (i.e., a balanced trial). Their properties were also compared by simulation and using data from a cluster randomization trial.

A major contribution of this thesis is the derivation of the kappa-type ICC estimators and evaluation of their small sample properties. Similar evaluations were conducted for Cohen's kappa and the ANOVA ICC estimator. Both spaced scores (i.e., 1,2,3) and midrank scores were considered to calculate the ANOVA and kappa-type estimators. The algebraic comparison was presented in Chapter 2. It was shown that the ANOVA and kappa-type ICC estimators were asymptotically equivalent in a balanced trial as the number of clusters becomes large. Simulation results showed that kappa-type estimators were more close to the true values than ICC estimators when cluster sizes were fixed and small for $\rho = 0.005$ or $\rho = 0.01$. Conversely, ANOVA ICCs had relatively smaller bias in the case of variable cluster sizes. In addition, midrank scores reduced the biases of both kappa and ANOVA ICC estimators for $\rho = 0.005$ or $\rho = 0.01$ when cluster sizes are variable and small (i.e., $\mu = 50$).

Another contribution of this thesis is the derivation of the adjusted Cochran-Armitage test statistics obtained by directly applying simple correction terms accounting for clustering. The algebraic comparisons in Chapter 3 show that the three adjusted statistics are identical in a balanced trial. Simulation results indicated that statistics using both kappa-type and ANOVA ICC estimators generated satisfactory type I error rates at the 5% nominal level when cluster sizes were fixed. When cluster sizes were variable, however, the adjusted statistics using ANOVA ICC estimators resulted in satisfactory type I error rates under most parameter combinations. Among the tests which have valid type I error rates, the statistical power of the WLS C-A test using midrank ANOVA estimates (i.e., χ'^2_{A3}) was slightly higher than that of other test statistics while the difference not more than 2%. One possible reason may be the WLS approach yielded the more precise parameter estimates than the first two adjusted statistics which used the OLS approach.

Finally, the small-sample performance of GEE robust tests were improved by the adjustment approaches derived in Chapter 5. A total of sixteen model-based test statistics were compared in the simulation study. For fixed cluster sizes, all test statistics, except GEE model-based and robust Wald statistics, showed generally satisfactory type I error rates at the 5% nominal level. However, for variable cluster sizes, only the robust score test and the adjustment methods W_{BC1} , $W_{BC4}$and W_{df5} were shown to maintain the overall satisfactory type I error rates. Among the methods that resulted in valid type I error rates, the adjusted method W_{BC2}yielded the highest statistical power for fixed cluster sizes, and W_{BC1}yielded the highest power for variable cluster sizes.

9.1.2 Recommendations and Discussions

Our results indicate that adjusted Cochran-Armitage tests are reasonable choices for testing the intervention effect for ordinal outcome data obtained from cluster randomization trials when there are no complex analyses required (e.g., analysis of covariates). In particular, the WLS adjusted C-A test obtained using the midrank ANOVA ICC estimator performs best, especially for variable cluster sizes, in terms of type I error and power.

Small-sample performance of GEE robust Wald tests are seen to be improved by using adjustments and corrections. In particular, the adjusted test W_{BC1} is the most appropriate method in terms of type I error and power.

In contrast to the liberal behaviour of the GEE robust Wald test, the GEE robust score test tends to have a smaller test size than the nominal level. However, our simulation and example study results are not consistent with this discussion of the conservative behavior of robust score test. In particular, the robust score test statistic S_R yields satisfactory type I error rates under all parameter combinations in our simulation study. Also, our example study showed the p-value generated by S_R (i.e., 0.8610) is very similar to the p-value the GEE robust Wald test statistic generated (i.e., 0.8606). The above discrepancy between the discussions in Chapter 5 and our study results may be explained by the fact that the total number of clusters in our study is close to 30. According to Guo et al. (2005)'s research, the type I error rate of robust score tests approaches 0.05 as the number of clusters from two groups increases to 30.

In addition, we discussed in Chapter 4 that the regression coefficient estimate from marginal models is smaller than that from cluster-specific models. However, this relationship is seen in our simulation study to hold only for the parameter combinations where the log odds ratio θ is set to 1.2. One possible reason is that the regression coefficient estimates (log odds ratio) from both marginal and cluster-specific models are close to zero in both the example data and the simulation study with $\theta = 1$. Combined with the small ICC, this may explain why the two coefficient estimates are very similar to each other.

9.2 Limitations and Future Research

First, a potential topic for future research is to unify different methods of analysis of clustered ordinal outcomes data. For instance, the model-based tests are often equivalent, at least in special cases, to well known non-parametric test statistics. The challenge is that some adjustment for these tests will be needed when applied to clustered ordinal data.

In particular, the Cochran-Armitage test statistic is equivalent to the score statistic obtained from logistic regression analyses with an ordinal covariate (Cox, 1958). The Wilcoxon rank sum test when applied to compare two multinomial distributions with ordered categories is equivalent to the score test for proportional odds models using a binary covariate (McCullagh, 1980). Moreover, the two approaches are equivalent when the scores in the Cochran-Armitage trend test are set equal to the midrank for each group, as defined in the Wilcoxon rank sum test (Rosner, 2000; pp401). As such the Cochran-Armitage trend test unifies different methods that have been proposed to analyze independent ordinal data.

For clustered ordinal data, Jung and Kang (2001) derived a test statistic unifying the Wilcoxon rank sum test and the Cochran-Armitage trend test. In addition, Natarajan et al. (2012) formulated an estimating equations score test from the proportional odds model as an extension of the Wilcoxon rank sum test. As such, Jung and Kang's (2001) method could similarly unify different methods for the analysis of clustered ordinal outcome data as the Cochran-Armitage trend test does for the analysis of independent ordinal outcome data. In future research, Jung and Kang's method could be further explored to connect methods for the analysis of clustered ordinal outcome data.

Second, we have focused on methods that may be applied to the completely randomized design in this research. Although the extensions of these methods to stratified cluster randomization trials is fairly straightforward, the challenge of extending the methods to pair-matched designs poses problems that are an area for future research (Klar and Donner, 1997). One approach would be to break the matches for the design-based matching and apply the methods discussed above. Detailed evaluation of this approach, including the loss in power if the matching is effective, is needed.

Third, the approaches presented here were developed specially for the case of one intervention group and one control group. However, many trials contain more than two intervention groups. For example, the TVFSP data in our example originally had four intervention groups. The methods presented here may usefully be extended to trials with more than two intervention groups.

Fourth, the simulation study has only considered data with equal numbers of clusters per intervention group. This design restriction was made in order to understand the performance of the methods in simple scenarios. However, there is often considerable variation in the number of clusters in practice. An equal number of clusters per intervention group generally leads to an increase in efficiency as compared to unequal allocation (Donner and Klar, 2000, p.59). Further research is required to assess our findings to more general settings such as studies having unbalanced cluster numbers.

Fifth, we deliberately focused on community intervention trials, which typically enrol a small number of large clusters. This focus reflects the relatively greater methodological challenge of statistical inferences arising in these studies. For example, the validity of statistical inferences is often problematic when there are few large clusters. Therefore, as Koepsell et al. (1991, 1992) and Donner and Klar (2000, p100) suggested, particular care must be taken when applying methods requiring a large number of clusters (e.g., GEE using robust variance estimators) to community intervention trials. Conversely, the methods discussed in this thesis could be naturally applied to trials having a large number of small clusters, for example, for example, families.

Finally, the simulation study evaluating marginal and cluster-specific extensions of ordinal logistic regression models is limited to models with cumulative logit links. Although the most popular model for ordinal responses uses logits of cumulative probabilities (Lui and Agresti, 2005), other types of links (e.g., adjacent-category logits or continuation-ratio logits) may also be of interest for ordinal data analysis. Therefore further study may be helpful to broaden our findings to these other ordinal response regression models.

References

Abraira, V., and De Vargas, A. P. (1999), 'Generalization of the Kappa coefficient for ordinal categorical data, multiple observers and incomplete designs', *Qüestiió (Barcelona)* **23**, 561-571.

Agresti, A. (2002), *Categorical Data Analysis*, Wiley, New York.

Agresti, A. (2010), *Analysis of Ordinal Categorical Data*, Wiley, New Jersey.

Agresti, A., and Coull, B. A. (2002), 'The analysis of contingency tables under inequality constraints', *Journal of Statistical Planning and Inference* **107**, 45-73.

Agresti, A., and Natarajan, R. (2001), 'Modeling clustered ordinal categorical data: A survey', *International Statistical Review* **69**, 345-371.

Agresti, A., Lipsitz, S. R., and Lang, J. B. (1991), 'Analysis of sparse repeated categorical measurement data', *Proceeding of the 16th annual SAS user's group international conference, Cary, North Carolina, SAS Institute Inc,* 1452-1460.

Aitchison, J., and Silvey, S. D. (1958), 'Maximum-likelihood estimation of parameters subject to restraints', *Annals of Mathematical Statistics* **29**, 813-828.

Aitkin, M. (1999), 'A general maximum likelihood analysis of variance components in generalized linear models', *Biometrics* **55**,117-128.

Armitage, P. (1955), 'Tests for linear trends in proportions and frequencies', *Biometrics* **11**, 375-386.

Banerjee, M., Capozzoli, M., McSweeney, L., and Sinha, D. (1999), 'Beyond kappa: a review of interrater agreement measures', *Canadian Journal of Statistics* **27**, 3-23.

Bellamy, S. L., Li, Y., Lin, X., and Ryan, L. M., (2005), 'Quantifying PQL bias in estimating cluster-level covariate effects in generalized linear mixed models for group-randomized trials', *Statistica Sinica* **15**, 1015-1032.

Bellamy, S. L., Gibberd, R., Hancock, L., Howley, P., Kennedy, B., Klar, N., Lipsitz, S., and Ryan, L. (2000), 'Analysis of dichotomous outcome data for community intervention studies', *Statistical Methods in Medical Research* **9**, 135-159.

Biswas, A. (2004), 'Generating correlated ordinal categorical random samples', *Statistics and Probability Letters* **70**(1), 25-35.

Bland, M. (2000), *An Introduction to Medical Statistics*, Oxford University Press, USA.

Boos, D. D. (1992), 'On generalized score tests', *The American Statistician* **46**,327-333.

Booth, J. G., and Hobert, J. P. (1999), 'Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm', *Journal of the Royal Statistical Society Series B (Statistical Methodology)* **61**, 265-285.

Bradley, J. V. (1978), 'Robustness?', *British Journal of Mathematical and Statistical Psychology* **31**, 144-152.

Brant, R. (1990), 'Assessing proportionality in the proportional odds model for ordinal logistic regression', *Biometrics* **46**, 1171-1178.

Brenner, H., and Kliebsch, U. (1996), 'Dependence of weighted kappa coefficients on the number of categories', *Epidemiology* **7**(2) 192-202.

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

Brier, S. (1980). 'Analysis of contingency tables under cluster sampling', *Biometrika* **67**, 591-560.

Brunner, E., and Langer, F. (2000), 'Nonparametric analysis of ordered categorical data in designs with longitudinal observations and small sizes', *Biometrical Journal* **42**, 663-675.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006), 'The design of simulation studies in medical statistics', *Statistics in Medicine* **25**(24), 4279-92.

Byng, R., Jones R., Leese M., Hamilton B., McCrone P., and Crai, T. (2004), 'Exploratory cluster randomised controlled trial of shared care development for long-term mental illness', *The British Journal of General Practice* **54**, 259–266.

Campbell, M. J., Fayers, P. M., and Grimshaw, J. M. (2005), 'Determinations of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research', *Clinical Trials* **2**, 99-107.

Chen, J. J., and Li. L. A. (1994), 'Dose-response modeling of trinomial responses from developmental experiments', *Statistica Sinica* **4**, 265–274.

Chuang, C., and Cox, C. (1985), 'Pseudo maximum likelihood estimation for the Dirichlet-Multinomial distribution', *Communications in Statistics (Theory and Methods)* **14**(10), 2293-2311.

Cochran, W. G. (1954), 'Some methods for strengthening the common X2 tests', *Biometrics* **10**, 417-451.

Cochan, W. G. (1997), *Sampling Techniques, Third Edition,* John Wiley & Sons, Inc., New York.

Cohen, J. (1960), 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement* **20**, 37-46.

Cohen, J. (1968), 'Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit', *Psychological Bulletin* **70**(4), 213–220.

Cook, T. D., and DeMets, D. L. (2008), *Introduction to Statistical Methods for Clinical Trials,* Chapman & Hall/CRC, Boca Raton, Florida.

Coull, B. A., and Agresti, A. (2000), 'Random effects modeling of multiple binomial response using the multivariate binomial logit-normal distribution', *Biometrics* **56**, 73-80.

Cox, D. R. (1958), 'The regression analysis of binary sequences', *Journal of Royal Statistical Society, Series B* **20**, 215-242.

Demirtas, H. (2006), 'A method for multivariate ordinal data generation given marginal distributions and correlations', *Journal of Statistical Computation and Simulation* **76**, 1017-1025.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

Donner A., (1986), 'A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model', *International Statistical Review* **54**, 67-82.

Donner, A., and Banting, D. (1988), 'Adjustment of frequently used chi-square procedures for the effects of site-to-site dependence in the analysis of dental data', *Journal of Dental Research* **68**, 1350-1354.

Donner, A., and Donald , A. (1987), 'Analysis of data arising from a stratified design with the cluster as unit of randomization', *Statistics in Medicine* **6**(1), 43-52.

Donner, A., and Donald, A. (1988), 'The statistical analysis of multiple binary measurements', *Journal of Clinical Epidemiology* **41**, 899-905.

Donner A., and Klar N. (1996), 'Statistical considerations in the design and analysis of community intervention trials', *Journal of Clinical Epidemiology* **49**(4), 435-9.

Donner, A., and Klar, N. (2000), *Design and Analysis of Cluster Randomization Trials in Health Research,* Edward Arnold Publishers Ltd.

Donner, A., and Klar, N. (2004), 'Pit falls of and controversies in cluster randomization trails', *American Journal of Public Health* **94**, 416-422.

Donner, A., and Koval, J. J. (1987), 'A procedure for generating groups sizes from a one-way classification with a specified degree of imbalance', *Biometrical Journal of Mathematical Methods in Bioscience* **20**,181-187.

Donner, A., Eliasziw, M., and Klar, N. (1994), 'A comparison of methods for testing homogeneity of proportions in teratologic studies', *Statistics in Medicine* **13**, 1253-1264.

Eldridge S. M., Ashby D., and Kerry S. (2006), 'Sample size randomized trials: effect of coefficient of variation of cluster size and analysis method', *International Journal of Epidemiology* **35**, 1292-1300.

Elston, R. (1977), 'Response to query: estimating "heritability" of a continuous trait', *Biometrics* **33**, 232-233.

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 Computing Simulation* **4**, 19-29.

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

Feinstein, A. R., and Cicchettti, D. V. (1990), 'High agreement but low kappa', *Journal of Clinical Epidemiology* **43**, 543-549.

Feng, Z. D., and Braun, T. M. (2002), 'Small sample inference for clustered data', *Proceedings of the Seattle Symposium in Biostatistics*.

Feng, Z., McLerran, D., and Grizzle, J. (1996), 'A comparison of statistical methods for clustered data analysis with Gaussian error', *Statistics in Medicine* **15**, 1793-1806.

Fielding, A. (1999), 'Why use arbitrary points score? Ordered categories in models of educational progress', *Journal of the Royal Statistical Society, Series A*. **162**, 303-328.

Fielding, A., Yang, M., and Goldstein, H. (2003), 'Multilevel ordinal models for examination grades', *Statistical Modeling* **3**, 127-153.

Firth, D. (1993), 'Bias reduction of maximum likelihood estimates', *Biometrika* **80**, 27–38.

Fitzmaurice, G. M., and Laird, N. M. (1993), 'A likelihood-based method for analysing longitudinal binary responses', *Biometrika* **80**, 141-152.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004), *Applied Longitudinal Analysis*. Wiley, New Jersey.

Flay, B. R., Miller, T. Q., Hedeker, D., Siddiqui, O., Britton, C. F., Brannon, B. R., Johnson, A., Hansen, W. B., Sussman, S., and Dent, C. (1995), 'The Television, school, and family smoking prevention and cessation project', *Preventive Medicine* **24**, 29-40.

Flay, B., Brannon, B., Johnson, C., Hansen, W., Ulene, A., Whitney,-Saltiel, D., Gleason, L., Sussman, S., Gavin, M., Kimarie, G., Sobol, D., and Spiegel, D. (1988), 'The television school and family smoking and cessation project: I. Theoretical basis and program development', *Preventive Medicine* **17**, 585-607.

Fleiss, J. L. (1981) *Statistical Methods for Rates and Proportions*. John Wiley, New York.

Fleiss, J. L., and Cohen, J. (1973), 'The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability', *Educationa and Psychological Measurement* **33**, 613-619.

Fleiss, J. L., and Cuzick, J. (1979), 'The reliability of dichotomous judgements: Unequal numbers of judges per subject', *Applied Psychological Measurement* **3**, 537-542.

Fung, K. Y., Krewski, D., Rao, J. N., and Scott, A. J. (1994), 'Tests for Trend in developmental toxicity trend Experiments with Correlated Binary Data', *Risk Analysis* **14**, 639-648.

Gail, M. H., Byar, D. P., Pechacek, T. F., and Corle, D. K. (1992), 'Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT)', *Controlled Clinical Trials* **13**, 6-21.

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

Gattellaria, M., Donnelly, N., Taylor, N., Meerkin, M., Hirst, G., and Ward, E. (2005), 'Does 'peer coaching' increase GP capacity to promote informed decision making about PSA screening? A cluster randomized trial', *Family Practice* **22**, 253-265.

Geys, H., Molenberghs, G., and Ryan, L. (1999), 'Pseudo-likelihood modeling Of multivariate outcomes in developmental toxicology', *Journal of the American Statistical Association*, **94**(447), 734–745.

Gibbons, R. D., and Hedeker, D. (1997), 'Random-effects probit and logistic regression models for three-level data', *Biometrics* **53**, 1527-1537.

Glonek, G. F. V. (1996), 'A class of regression models for multivariate categorical responses', *Biometrika* **83**, 15-28.

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

Goodman, L. A. (1985), 'The analysis of cross-classified data having ordered and unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries', *Annals of Statistics* **13**, 10-69.

Graubarbd., I., and Korne, L. (1987), 'Choice of column scores for testing independence in ordered 2 x K tables', *Biometries* **43**, 471-476.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969), 'Analysis of categorical data by linear models', *Biometrics* **25**, 489-504.

Guo, X., Pan, W., Connett, J. E., Hannan, P. J., and French, S. A. (2005), 'Small-sample performance of the robust score test and its modifications in generalized estimating equations', *Statistics in Medicine* **24**, 3479-3495.

Haber, M. (1985), 'Maximum likelihood methods for linear and log-linear models in categorical data', *Computational Statistics and Data Analysis* **3**, 1-10.

Haber, M., and Brown, M. B. (1986), 'Maximum likelihood methods for log-linear models when expected frequencies are subject to linear constraints', *Journal of the American Statistical Association* **81**, 477-482.

Hannan, P. J., Murray, D. M., Jacobks, D. J., and McGovern, P. G. (1994), 'Parameters to aid in the Design and Analysis of community trials: Intraclass Correlations from the Minnesota Heart Health Program', *Epidemiology* **5**, 88-95.

Hartzel, J., Agresti, A., and Caffo, B. (2001), 'Multinomial logit random effects models', *Statistical Modeling* **1**, 81-102.

Hauck, W. W., and Donner, A. (1977), ''Wald's test as applied to hypotheses in logit analysis', *Journal of the American Statistical Association* **72**(360), 851-853.

Howard-Pihney, B., Winkleby, M. A., Albright, C. I., Bruce, B., and Fortmann, S. P. (1997), 'The Stanford nutrition action program: a dietary fat intervention for low-loteracy adults', *American Journal of Public Health* **87**, 1971-1976.

Heagerty P. J., and Zeger, S. L. (2000), 'Marginalized Multilevel Models and Likelihood Inference', *Statistical Science* **15**(1), 1-26.

Hedeker, D. (2003), 'A mixed-effects multinomial logistic regression model', *Statistic in Medicine* **22**, 1433-1446.

Hedeker, D., and Gibbons, R. D. (1994), 'A random-effects ordinal regression model for multilevel analysis', *Biometrics* **50**, 933-944.

Hedeker, D., and Gibbons, R. D. (1996), 'MIXOR: A computer program for mixed-effects ordinal regression analysis', *Computer Methods and Programs in Biomedicine* **49**, 157-176.

Hedeker, D., and Mermelstein, R. J. (1998), 'A multilevel thresholds of change model for analysis of stages of change data', *Multivariate Behavioral Research* **33**, 427-455.

Hedeker, D., Gibbons, R. D., and Flay, B. R. (1994), 'Random-effects regression models for clustered data with an example from smoking prevention research', *Journal of Consulting and Clinical Psychology* **62**, 757-765.

Heo, M., and Leon, A. (2005), 'Comparison of multiplicity adjustment strategies for correlated binary endpoints', *Journal of Biopharmaceutical Statistics* **15**, 839-855

Hinkley, D. V. (1977), 'Jackknifing in unbalanced situations', *Technometrics* **19**, 282-292.

Hosmer, D., and Lemeshow, S. (2000), *Applied Logistic Regression,* Wiley, New York.

Imrey, P. B. et al (1981), 'Categorical data Analysis: Some reflections on the loglinear model and Logistic Regression, PART 1: Historical and Methodological Overview', *International Statistical Review* **49**, 265-283.

Jang, W., and Lim, J. (2006), '*PQL estimation biases in generalized linear mixed models*. Working paper 05-21', Institute of Statistics and Decision Sciences, Duke University, Durham, NC, USA.

Jung, S., and Kang, S. (2001), 'Tests for 2*K contingency tables with clustered ordered categorical data', *Statistics in Medicine* **20**, 785-794.

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.

Kim, H. Y., Williamson, J. M., and Lyles, C. M. (2005), 'Sample-size calculations for studies with correlated ordinal outcomes', *Statistics in Medicine* **24**, 2977-2987.

Kimeldorf, G., Sampson, A. R., and Whitaker, L. R. (1992), 'Min and max scoring for two-sample ordinal data', *Journal of the Americal Statistical Association* **87**, 241-247.

Kish, L. (1965), *Survey Sampling*, Wiley, New York.

Klar, N. (1993), *Stratified Cluster Randomization Trials*, Ph.D. Thesis, The University of Wetern Ontario.

Klar, N., and Donner, A. (1997), 'The merits of matching in community intervention trials: a cautionary tale', *Statistics in Medicine* **16**, 1753-1764.

Klepp, K.-I., Ndeki, S. S., Leshabari, M. T., Hannan, P. J., and Lyimo, B. A. (1997), 'AIDS Education in Tanzania: Promoting Risk Reduction among Primary School Children', *American Journal of Public Health* **87**, 1931-1936.

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, J., and Lehnen, R. (1977), 'A general methodology for the analysis of experiments with repeated measurement of categorical data', *Biometrics* **33**, 133-158.

Koepsell, T. D., Martin, D. C., Dichr P. H., Psaty, B. M., Wagner, E. H., Perrin, E. B., and Cheadle, A. (1991), 'Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs:a mixed-model analysis variance approach', *Journal of Clinical Epidemiology* **44**(7), 701-713.

Koepsell, T. D., Wagner, E. H., and Cheadle, A. C. et al. (1992), 'Selected methodological issues in evaluating community-based health promotion and disease prevention programs', *Annual Review of Public Health* **13**, 31-57.

Kupper, L. L., Portier, C., Hogan, M. D., and Yamamoto, E. (1986), 'The impact of litter effects on dose-response modeling in teratology', *Biometrics* **4**(2), 85-98.

Lang, J. B., and Agresti, A. (1994), 'Simultaneously modeling joint and marginal distributions of multivariate categorical responses', *Journal of the American Statistical Association* **89**, 625-632.

Leyland, A. H., and Goldstein, L. H. (2001), *Multilevel Modeling of Health Statistics*, Wiley, New York.

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

Lipsitz, S. R., and Ryan, L. M. (2000), 'Analysis of dichotomous outcome data for community intervention studies', *Statistical Methods in Medical Research* **2**, 135-160.

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

Lipsitz, S. R., Fitzmaurice, G. M., Orav, E. J., and Laird, N. M. (1994b), 'Performance of generalized estimating equations in practical situations', *Biometrics* **50**, 270-278.

Liu, I., and Agreti, A. (2005), 'The analysis of ordered categorical data: an overview and a survey of recent developments', *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* **14,** 1-73.

Liu, Q., and Pierce, D. A. (1994), 'A note on Gauss-Hermite quadrature', *Biometrika* **81**, 1602-1618.

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

Luepker, R. V., Perry, C. L., McKinlay, S. M., Nader, P. R., Parcel, G. S., Stone, E. J., Webber, L. S., Elder, J. P., Feldman, H. A., Johnson, C. C, et al. (1996),

'Outcomes of a field trial to improve children's dietary patterns and physical activity. The Child and Adolescent Trial for Cardiovascular Health. CATCH collaborative group', *Journal of the American Medical Association* **275**(10), 768-776.

Lui, K. J. (2002), 'Notes on estimation of the general odds ratio and the general risk Biometrical difference for paired-sample data', *Biometrical Journal* **44**(8), 957-968.

Lui, K. J., Cumberland, W. G., Mayer, J. A., and Eckhardt, L. (1999), 'Interval Estimation for the intraclass correlation in Dirichlet-Multinomial data', *Psychometrika* **64**, 355-369.

Lumley, T. (1996), 'Generalized estimating equations for ordinal data: a note on working correlation structures', *Biometrics* **52**, 354-361.

MacKinnon, J. G., and White, H. (1985), 'Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties', *Journal of Econometrics* **29**, 305-325.

Maclure, M., and Willett, W. C. (1987), 'Misinterpretation and misuse of the kappa statistic', *American Journal of Epidemiology* **126**, 161-169.

Mak, T. K. (1988), 'Analysing intraclass correlation for dichotomous variables', *Applied Statistics* **37**, 344-352.

Mancl, L. A., and DeRouen, T. A. (2001), 'A covariance estimator for GEE with improved small-sample properties', *Biometrics* **57**, 126 – 134.

Mantel, N., and Haenszel, W. (1959), 'Statistical aspects of the analysis of data from retrospective studies of disease', *Journal of the National Cancer Institute* **22**, 719-748.

Marinacci, C., Schifano. P., Borgia. P., and Perucci. C. A. (2001), 'Application of random effect regression model for outcome evaluation of two controlled trials', *Statistics in Medicine* **20**, 3769-3776.

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

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models,* Chapman and Hall, London.

McCulloch, C. E. (1997), 'Maximum likelihood algorithms for generalized linear mixed models', *Journal of American Statistical Association* **92**, 162-170.

McCusker, J., Stoddard, A. M., ScD, Zapka, J. G., Morrison, C. S., Zorn, M., and Lewis, B. F. (1992), 'AIDS education for drug abusers: evaluation of short-term effectiveness', *American Journal of Public Health* **82**, 533-540.

McKelvey, R. D., and Zavoina, W. (1975), 'A statistical model for the analysis of ordinal level dependent variables', *Journal of Mathematical Sociology* **4**, 103-120.

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.

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, New York.

Morel, J., and Nagaraj, N. K. (1993), 'A finite mixture distribution for modeling multinomial extra variation', *Biometrika* **80**, 363-372.

Murray, D. M. (1998), *Design and Analysis of Group-Randomized Trials,* Oxford University Press, USA.

Murray, D. M., Varnell, S. P., and Blistein, J. L. (2004), 'Design and analysis of grouped-randomized trials: A review of recent methodological developments', *American Journal of Public Health* **94**, 423-432.

Murray, D. M., Pals, S. L., Blitstein, J. L., Alfano, C. M., and Lehman, J. (2008), 'Design and Analysis of Group-Randomized Trials in Cancer: A Review of Current Practices', *Journal of the National Cancer Institute* **100**, 483-491.

Murray, D. M., Perry, C. L., Griffin, G., Harty, K. C., Jacobs, D. R. Jr, Schmid, L., Daly, K., and Pallonen, U. (1992), 'Results from a statewide approach to adolescent tobacco use prevention', *Preventive Medicine* **21**, 449-72.

Neuhaus, J. M., and Segal, M. R. (1993), 'Design effects for binary regression models fitted to dependent data', *Statistics in Medicine* 12(13), 1259-1268.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991), 'A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data', *International Statistical Review* **59**, 25 –35.

Narayanan, A. (1991), 'Maximum likelihood estimation of the parameters of dirichlet distribution', *Applied Statistics* **40**(2), 365-374.

Natarajan, S., Lipsitz, S. R., Sinha, D., and Fitzmaurice, G. (2012), 'An extension of the Wilcoxon rank-sum test for complex sample survey data', *Applied Statistics* **61**, 653-664.

Pan, W. (2001), 'On the robust variance estimator in generalised estimating equations', *Biometrika* **88**, 901-906.

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.

Parsons, N. R., Costa, M. L., Achten, J., and Stallard, N. (2009), 'Repeated meandurements proportional odds logistic regression analysis of ordinal score data in the statistical software package R', *Computational Statistics and Data Analysis* **53**, 632-641.

Patton, G. C., Bond, L., Carlin, J. B., Thomas, L., Butler, H., Glover, S., Catalano, R., and Bowes, G. (2006), 'Promoting Social Inclusion in Schools: A Group-Randomized Trial of Effects on Student Health Risk Behavior and Well-Being', *American Journal of Public Health* **96**, 1582-1587.

Paul, S. R., Balasooriya, U., and Banerjee, T. (2005), 'Fisher Information Matrix of the Dirichlet-multinomial Distribution', *Biometrical Journal* 47(2), 230-236.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R. (1996), 'A survey of methods for analyzing correlated binary response data', *International Statistical Review* **64**, 89-118.

Peterson, B., and Harrell, F. E. Jr. (1990), 'Partial proportional odds models for ordinal response variables', *Applied Statistics* **39**, 205-217.

Pinheiro, J. C., and Chao, E. C. (2006), 'Efficient LaPlacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models', *Journal of Computational and Graphical Statistics* **15**, 58–81.

Preisser, J. S., and Qaqish, B. F. (1996), 'Deletion diagnoistics for generalized estimating equations', *Biometrika* **83**, 551-562.

Rabe-Hesketh, S., and Skrondal, A. (2002), 'Reliable estimation of generalized linear mixed models using adaptive quadrature', *The Stata Journal* **2**, 1-21.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2005), 'Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects', *Journal of Econometrics* **128**, 301-323.

Raman, R., and Hedeker, D. (2005), 'A mixed-effects regression model for three-level ordinal response data', *Statistics in Medicine* **24**, 3331-3345.

Rao, J. N. K., and Scott, A. J. (1992), 'A simple method for the analysis of clustered binary data', *Biometrics* **48**, 77-585.

Rao, J. N. K., and Thomas, D. R. (1988), 'The analysis of cross-classified categorical data from complex sample surveys', *Sociological Methodology* **18**, 213-269.

Ridout, M. S., Demetrio, C. G. B., and Firth, D. (1999), 'Estimating intraclass correlation for binary data', *Biometrics* **55**, 137-148.

Rodriguez, G., and Goldman, N. (2001), 'Improved estimation procedures for multilevel models with binary response: a case study', *Journal of the Royal Statistical Society, Series A*. **164**, 339–355.

Rosendal, M., Bro, F., Fink, P., Christensen, K. S., and Olesen, F. (2003), 'Diagnosis of somatisation: effect of an educational intervention in a cluster randomised controlled trial', *British Journal of General Practice* **53**, 917–922.

Rosner, B. (2000), *Fundamentals of Biostatistics,* Duxbury Press, Boston.

Rosner, B., and Grove, D. (1999), 'Use of the Mann-Whitney-U-test for clustered data', *Statistics in Medicine* **18**, 1387-1400.

Rosner, B., Glynn, R. J., and Lee, M. L. (2003), 'Incorporation of clustering effects fort he Wilcoxon rank sum test: A large sample approach', *Biometrics* **59**, 1089-1098.

Rosner, B., Glynn, R. J., and Lee, M. L. (2006), 'The Wilcoxon signed rank test for paired comparisons of clustered data', *Biometrics* **62**, 185-192.

Rosner, B., Glynn, R. J., and Lee, M. L. (2006), 'Extention of the ranked sum test for clustered data: two-group comparisons with group membership defined at the subunit level', *Biometrics* **62**, 1251-1259.

Rotnitzky, A., and Jewell, N. P. (1990), 'Hypothesis Testing of Regression Parameters in Semi-Parametric Generalized Linear Models for Cluster Correlated Data', *Biometrika* **77**, 485-497.

Qu, Y., Piedmonte, M. R., and Medendorp, S. V. (1995), 'Latent variable models for clustered ordinal data', *Biometric* **51**, 268-275.

Qu, Y., Piedmonte, M. R., and Williams, G. W. (1994), 'Small sample validity of latent variable models for correlated binary data', *Communications in Statistics: Simulations* **23**, 243-269.

Santos, D. M., and Berridge, D. M. (2000), 'A continuation ratio random effects model for repeated ordinal responses', *Statistics in Medicine* **19**, 3377-3388.

Sashegyi, A., Brown, S., and Farrell, P. (2000), 'Application of generalized random effects regression models for cluster-correlated longitudinal data to a school-based smoking prevention trial', *American Journal of Epidemiology* **152**, 1192–1200.

Scatterthwaite, F. F. (1941), 'Synthesis of variance', *Psychometrika* **6**, 309-316.

Schaeffer, N. C., and Presser, S. (2003), 'The Science of Asking Questions', *Annual Review of Sociology* **29**, 65-88.

Schnell, D. J., Magee, E., and Sheridan, J. R. (1995), 'A regression method for analyzing ordinal data from intervention trials', *Statistics in Medicine* **14**, 1177-1189.

Scott, W. A. (1955), 'Reliabuility of content analysis: the case of nominal scale coding', *Public Opinion Quart* **9**, 321-325.

Scott, A. J., and Holt, D. (1982), 'The effect of two-stage sampling on ordinary least squares methods', *Journal of the American Statistical Association* **77**, 848-854.

Seligman, H. K., Wang, F. F., Palacios, J. L., Wilson, C. C., Daher, C., Piette, J. D., and Schillinger, D. (2005), 'Physician notification of their diabetes patients' limited health literacy: A randomized, controlled trial', *Journal of General Internal Medicine* **20**, 1001-1007.

She, D., Li, Y., Zhang, H., Graubard B., I., and Li, Z. (2010), 'Trend tests for genetic association using population-based cross-sectional complex survey data', *Biostatistics* **11**(1), 48-56.

Sherman, M., and Le Cessie, S. (1997), 'A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear model', *Communications in Statistics: Simulation* **26**, 901-925.

Simpsom, J. M., Klar, N., and Donner, A. (1995), 'Accounting for cluster randomization: a review of primary prevention trails, 1990 through 1993', *American Journal of Public Health* **85**, 1378-1383.

Song, W. Y., Lee, Y. J., Hwang, S. W., Kim, H. Y., Yoo, B. H., and Kwon, O. J. (1997), 'Comparative study of tracheal anastomotic techniques', *Korean Journal of Thoracic and Cardiovascular Surgery* **30**, 1-7.

Stefanescu, C., and Turnbull, B. W. (2003), 'Likelihood inference for exchangeable binary data with varying cluster sizes', *Biometrics* **59**(1), 18-24.

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**(11), 1419–1433.

Stiger, T. R., Kosinski, A. S., Barnhart, H. X., and Kleinbaum D. G. (1998), 'ANOVA for repeated ordinal data with small sample size? A comparison of ANOVA, MANOVA, WLS and GEE methods by simulation', *Communications in Statistics – Simulation and Computation* **27**, 357-375.

Streiner, D. L., and Norman, G. R. (2008), *Health Measurement Scales: A Practical Guide to Their Development and Use*, Oxford University Press, USA

Stroud, H., and Sechrest, D. (1966), *Gaussian Quadrature Formulas*, Prentice Hall, Englewood Cliffs, N.J.

Sullivan, L. M., and D'Agostino, Sr. R. B. (2003), 'Robustness and power of analysis of covariance applied to ordinal scaled data as arising in randomized controlled trials', *Statistics in Medicine* **22**, 1317-1334.

Tamura, R. N., and Young, S. S. (1986), 'The incorporation of Historical control information intests of proportions: Simulation study of Tarone's procedure', *Biometrics* **42**, 343-349

Tamura, R. N., and Young, S. S. (1987), 'A stabilized moment estimator for the beta-binomial distribution', *Biometrics* **43**, 813-824.

Ten Have T. R., Landis, J. R., and Hartzel, J. (1996), 'Population-averaged and cluster-specific models for clustered ordinal response data', *Statistics in Medicine* **15**, 2573-2588.

Tsou, T. S., and Shen, C. W. (2008), 'Parametric robust inferences for correlated ordinal data', *Statistics in Medicine* **27**(18), 3550–3562.

Tutz, G., and Hennevogl, W. (1996), 'Random effects in ordinal regression models', *Computing Statistical Data Analysis* **22**, 537-557.

Ukoumunne, O. C., and Thompson, S. G. (2001), 'Analysis of cluster randomized trials with repeated cross-sectional binary measurements', *Statistics in Medicine* **20**, 417-433.

Varnell, S. P., Murray, D. M., Janega, J. B., and Blitstein, J. L. (2004), 'Design and analysis of group-randomized trails: a review of recent practices', *American Journal of Public Health* **94**, 393-399.

Vermunt, J., and Hagenaars, J. A. (2004), 'Ordinal longitudinal data analysis', *Methods in human growth research,* 374-393, Cambridge University Press, UK.

Villar, J., Ba'aqeel, H., Piaggio, G., Lumbiganon, P., Miguel, Belizán, J., Farnot, U., Al-Mazrou, Y., Carroli, G., Pinol, A., Donner, A., Langer, A., Nigenda, G., Mugford, M., Fox-Rushby, J., Hutton, G., Bergsjø, P., Bakketeig, L., and Berendes, H., (2001), 'WHO antenatal care randomised trial for the evaluation of a new model of routine antenatal care', *Lancet* **357**, 1551-64.

Watson, M., Kendrick, D., Coupland, C., Woods, A., Futers, D., and Robinson, J. (2005), **'**Providing child safety equipment to prevent injuries: randomised controlled trial', *British Medical Journal* **330**(7484), 178.

Yamamoto, E., and Yanagimoto T. (1992), 'Moment estimators for the beta-binomial distribution', *Journal of applied statistics* **33**, 273-283.

Yang M. (2001), 'Multinomial regression', *Multilevel Modelling of Health Sciences*, 107–125.

Yates, F. (1948), 'The analysis of contingency tables with groups based on quantitative characters', *Biometrika* **35**, 176-181.

Zeger, S., and Karim, R. (1991), 'Generalized linear models with random effects', *Journal of American Statistical Association* **86**, 79-86.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988), 'Models for longitudinal data: a generalized estimating equation approach', *Biometrics* **44**(4), 1049-60.

Zhang, L. (2009), *Risk Factor Analyses in Matched-Pair Cluster Randomization Trials*, M.Sc. Thesis, The University of Western Ontario.

Zou, G. Y. (2002), *Interim Analysis for Cluster Randomization Trials with Binary Outcomes*, Ph.D. Thesis, The University of Wetern Ontario.

Zou, G. Y., Donner, A., and Klar, N. (2005), 'Group sequential methods for cluster randomization trials with binary outcomes', *Clinical Trials* **2**(6), 479-487.

Zucker, D. M., Lakatos, E., Webber, L. S. et al. for the CATCH Study Group. (1995), 'Statistical design of the child and Adolescent trial for cardiovascular health (CATCH): Implications of cluster randomization', *Controlled Clinical Trials* **16**, 96-118.

Appendix A

Matrix version derivation of weighted least squares Cochran-Armitage estimation

The linear probability model used to test the trend for clustered binary outcome data is written as $E(\hat{P}) = \alpha_C + \beta_C S$. We denote P_i as a $(\sum_{j=1}^{n_i} m_{ij}) \times 1$ outcome vector and S_i as a $2 \times \sum_{j=1}^{n_i} m_{ij}$ score matrix. Then $P = \left[P_1^{'}, P_2^{'}, ...P_G^{'}\right]^{'}$ and $S = \left[S_1^{'}, S_2^{'}, ...S_G^{'}\right]^{'}$. By minimizing the weighted square of the error terms

$$\sum_{i=1}^{G} W_i (P_i - \alpha_C - \beta_C \times S_i)^2$$

the WLS estimator of β_c is given by

$$\hat{\beta}_C = (S'WS)^{-1} S'WP$$

where W was defined in Section 3.3.3.

Letting $c_{ij} = 1 + (m_{ij} - 1)\rho$, then

$$[S'WS]^{-1} = \begin{bmatrix} \sum_{i=1}^{G} \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} & \sum_{i=1}^{G} S_i \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \\ \sum_{i=1}^{G} S_i \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} & \sum_{i=1}^{G} S_i^2 \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} \sum_{i=1}^{G} S_i^2 \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} & -\sum_{i=1}^{G} S_i \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \\ -\sum_{i=1}^{G} S_i \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} & \sum_{i=1}^{G} \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \end{bmatrix}$$

and

$$S'WP = \begin{bmatrix} \sum_{i=1}^{G} \hat{P}_i \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \\ \sum_{i=1}^{G} S_i \hat{P}_i \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \end{bmatrix}.$$

So the numerator of $\hat{\beta}_{CB-WLS}$ is given by

$$\sum_{i=1}^{G} \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \left(\sum_{i=1}^{G} S_i \hat{P}_i \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} - \tilde{P}\tilde{S} \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \right)$$

$$= \sum_{i=1}^{G} \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \left(\sum_{i=1}^{G} (S_i - \tilde{S})(\hat{P}_i - \tilde{P}) \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \right).$$

The denominator of $\hat{\beta}_{CB-WLS}$ is given by

$$\sum_{i=1}^{G} \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \left(\sum_{i=1}^{G} S_i^2 \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} - \tilde{S}^2 \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \right)$$

$$= \sum_{i=1}^{G} \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \left(\sum_{i=1}^{G} (S_i - \tilde{S})^2 \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \right).$$

Thus the WLS estimator is given by

$$\hat{\beta}_{CB-WLS} = \left(\sum_{i=1}^{G} (S_i - \tilde{S})(P_i - \tilde{P}) \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \right) \Big/ \left(\sum_{i=1}^{G} (S_i - \tilde{S})^2 \sum_{j=1}^{n_i} \frac{m_{ij}}{c_{ij}} \right).$$

Curriculum Vitae

Name: Ruochu Gao

Post-secondary Education and Degrees:

Department of Economics and Management
China University of Petroleum, Beijing, China
1996-2000 B.Sc. Economics

Department of Mathematics and Statistics
McMaster University, Hamilton, Ontario
2002-2004 M.Sc. Statistics

Department of Epidemiology and Biostatistics
The University of Western Ontario, London, Ontario
2004-2012 Ph.D. Biostatistics

Related Work Experience

Consulting Statistician
Biostatistical Support Unit
The University of Western Ontario
London, Ontario, 2004-2012

Manager, Biostatistics
Sanofi Pasteur
Toronto, Ontario, 2010-2011

Statistician
Apotex Inc.
Toronto, Ontario, 2011-Present

Teaching Experience

Teaching Assistant
Department of Mathematics and Statistics
McMaster University
Hamilton, Ontario, 2002-2004

Teaching Assistant
Department of Epidemiology and Biostatistics
The University of Western Ontario
London, Ontario, 2005-2010