Analysis and Diagnostics of Categorical Variables with Multiple Outcomes

by

Thomas Falk Suesse

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Statistics.
Victoria University of Wellington
2009

Abstract

Surveys often contain qualitative variables for which respondents may select any number of the outcome categories. For instance, for the question "What type of contraceptive have you used?" with possible responses (oral, condom, lubricated condom, spermicide, and diaphragm), respondents would be instructed to select as many of the $J = 5$ outcomes as apply. This situation is known as *multiple responses* and outcomes are referred to as items. This thesis discusses several approaches to analysing such data.

For stratified multiple response data, we consider three ways of defining the common odds ratio, a summarising measure for the conditional association between a row variable and the multiple response variable, given a stratification variable. For each stratum, we define the odds ratio in terms of: 1 item and 2 rows, 2 items and 2 rows, and 2 items and 1 row. Then we consider two estimation approaches for the common odds ratio and its (co)variance estimators for these types of odds ratios. The model-based approach treats the J items as a J-dimensional binary response and then uses logit models directly for the marginal distribution of each item by applying the generalised estimating equation (GEE) (Liang and Zeger 1986) method. The non-model-based approach uses Mantel-Haenszel (MH) type estimators.

The model-based (or marginal model) approach is still applicable for more than two explanatory variables. Preisser and Qaqish (1996) proposed regression diagnostics for GEE. Another model fitting approach is the homogeneous linear

predictor model (HLP) based on maximum likelihood (ML) introduced by Lang (2005). We investigate deletion diagnostics as the Cook distance and DBETA for multiple response data using HLP models (Lang 2005), which have not been considered yet, and propose a simple "delete=replace" method as an alternative approach for deletion. Methods are compared with the GEE approach.

We also discuss the modelling of a repeated multiple response variable, a categorical variable for which subjects can select any number of categories on repeated occasions. Multiple responses have been considered in the literature by various authors; however, repeated multiple responses have not been considered yet. Approaches include the marginal model approach using the GEE and HLP methods, and generalised linear mixed models (GLMM). For the GEE method, we also consider possible correlation structures and propose a groupwise correlation estimation method yielding more efficient parameter estimates if the correlation structure is indeed different for different groups, which is confirmed by a simulation study.

Ordered categorical variables occur in many applications and can be seen as a special case of multiple responses. The proportional odds model, which uses logits of cumulative probabilities, is currently the most popular model. We consider two approaches focusing on the mis-specification of a covariate. The binary approach considers the proportional odds model as $J-1$ logistic regression models and applies the cumulative residual process introduced by Arbogast and Lin (2005) for logistic regression. The multivariate approach views the proportional odds model as a member of the class of multivariate generalised linear models (MGLM), where the response variable is a vector of indicator responses.

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible.

I would like to express my deep and sincere gratitude to my primary supervisor, senior lecturer Dr. Ivy Liu. When I contacted Victoria University she voluntarily offered me an interesting Ph.D. topic and her supervision. She has provided assistance in numerous ways: She helped me financially by employing me as a research assistant, she always had great research ideas and always provided me with excellent supervision over the course of study. My next thanks are to Dr. Dong Wang, my secondary supervisor, for offering his help on many occasions and his expertise in diagnostics.

I am also very grateful to the University of Wellington for awarding me with the VUW Postgraduate Scholarship.

Special thanks to all people from the Statistics and Operations Research (STOR) programme for their help and support and for giving me the opportunity to be a tutor. I would like to thank the many other people at Victoria from the faculty office, such as Celia Simpson, and the school office for helping me in many different and often unbureaucratic ways. Special thanks for their support and understanding for my partial absence due to sickness.

I also wish to express my gratitude to Drs. Shirley and Ken Pledger, Christopher Ball, and Richard Gyde from editwrite.co.nz for revising the English of my manuscript.

My gratitude also to the three examiners, Prof. Steve Haslett, A/Prof. Christopher Bilder and Dr. Richard Arnold, for their helpful comments to improve the thesis, but also for their advise on publishing thesis results and addressing future research.

Lastly, I wish to thank my family: My parents Drs. Herbert and Irene Süße, who bore, raised, taught me, and loved me, and my fiancé Bianca Knoch for her support and love.

To them I dedicate this thesis.

Contents

Chapter 1

Introduction

This introduction extensively reviews methods for categorical data with multiple outcomes. The first section (Sec. 1.1) reviews multiple response data analysis through various models. Section 1.2 focuses on the Mantel-Haenszel (MH) methods for stratified data. Then Section 1.3 gives a review of diagnostic methods focusing on deletion diagnostics, and Section 1.4 shows an overview of the proportional odds model. The last section (Sec. 1.5) provides an outline of the thesis.

1.1 Review: Various Modelling Strategies for Multiple Response Data Analysis

1.1.1 Multiple Response Data

Surveys often contain qualitative variables for which respondents may select any number out of J outcome categories. For instance, Bilder and Loughin (2002) presented data, where 239 sexually active college women were asked "What type of contraceptives have you used?". They could select any answer of the following:

A-oral, B-condom, C-lubricated condom, D-spermicide, and E-diaphragm. Categorical variables that summarise this type of data are called *pick any/J variables* or *multiple response variables*, where J is the number of outcome categories ($J = 5$ in this case) and "/" stands for "out of "(Coombs 1964). Each outcome category is referred to as an *item* (Agresti and Liu 1999).

A special case of a multiple response variable is a multinomial variable only allowing J mutually exclusive outcome categories. We can cross-classify the counts from a survey that contains a pick any/J variable, along with some explanatory variables, into a contingency table. Table 1.1 by Bilder and Loughin (2002) presents such a cross-classification for the 239 sexually active college women with a group variable ($r = 2$ levels, whether a subject had a prior history of urinary tract infection (UTI) or not) and a stratification variable ($K = 2$ levels, the age groups) forming a $2 \times 5 \times 2$ contingency table. In this table, subjects may be represented in more than one cell.

Table 1.1: The marginal UTI data

	Contraceptive					Total	Total
	A	B	C	D	E	responses	women
Age $\geq$ 24							
UTI							
No	18	9	8	7	0	42	24
Yes	8	9	2	3	2	24	14
Age $<$ 24							
UTI							
No	55	41	37	27	0	160	85
Yes	75	68	33	22	5	203	116

The multiple response vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})^T$ with J items can be considered as a J-dimensional binary response vector, where i stands for the ith subject. For item j, the response is either "the item is selected" ($y_{ij} = 1$) or "the item is not selected" ($y_{ij} = 0$). The 2^J possible outcomes for $\mathbf{y}_i$ can be summarised in a joint

table and the full joint distribution, specified by 2^J probabilities, is characterised by $2^J - 1$ parameters.

1.1.2 Marginal Modelling

One approach, called *marginal modelling*, is modelling each component μ_{ij} of the mean response vector $\boldsymbol{\mu}_i$ of the multiple response variable, that is, modelling the (univariate) marginal distributions of $\mathbf{y}_i$. For multiple response data, the mean response μ_{ij} is identical to the probability of a positive response π_{ij}. The linear predictor $\eta_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\beta}_j$ is connected to π_{ij} by link function g_j such that $g_j(\pi_{ij}) = \eta_{ij}$, where $\boldsymbol{\beta}_j$ is the column vector of model parameters for the jth item and $\mathbf{z}_{ij}^T$ is the ith contribution to the design matrix of the jth model.

The joint model containing all J models can also be written in vector form as $\mathbf{g}(\boldsymbol{\pi}_i) = \mathbf{Z}_i \boldsymbol{\beta}$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_J^T)^T$, $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iJ})^T$ and $\mathbf{g} = (g_1, \ldots, g_J)^T$. The column vectors $\mathbf{z}_{ij}$ $(j = 1, \ldots, J)$ form matrix $\mathbf{Z}_i = \text{Diag}(\mathbf{z}_{i1}^T, \ldots, \mathbf{z}_{iJ}^T)$. We assume $\mathbf{Z}_i$ is an appropriate function of the covariate column vector $\mathbf{x}_i$. The convenient notation $\mathbf{g}(\boldsymbol{\pi}_i)$ stands for the column vector $(g_1(\pi_{i1}), \ldots, g_J(\pi_{iJ}))^T$.

For the UTI data, we could model the probability of a positive response for each item with a logistic link (g_j) and using the row variable ($r = 2$ levels) and the stratification variable ($K = 2$ levels) as explanatory variables. The most popular link is the logit link, which is also the canonical link for binary data. Other popular links are the probit, log-log and complementary log-log links (McCullagh and Nelder 1989). Marginal and other modelling strategies for multiple response data were presented by Agresti and Liu (1999), and Agresti and Liu (2001) among others.

The binary distribution is a member of the simple exponential family and maximum likelihood (ML) estimates can be easily obtained via the framework

of generalised linear models (GLM) (McCullagh and Nelder 1989). GLM were introduced by Nelder and Wedderburn (1972) although many models in the class of GLM were well established by then. The ML estimates of a GLM are the solutions of the likelihood equations (the derivatives of the log-likelihood), which only depend on the assumed distribution of the observations through the mean and variance. Within the class of GLM, the distribution determines the mean-variance relationship. Wedderburn (1974) introduced quasi-likelihood functions, where only an assumption about the mean-variance relationship is made without specifying the underlying distribution. For a GLM, the quasi-likelihood equations are identical to the likelihood equations. Both likelihood and quasi-likelihood equations are also often referred to as *score functions*.

However, the J dimensional binary vector $\mathbf{y}_i$ contains dependent observations. There are 2^J possible outcomes for $\mathbf{y}_i$. The underlying joint distribution is assumed to be multinomial and characterised by 2^J probabilities or $2^J - 1$ parameters. Treating the items naively as independent and applying ML estimation for each of the J models separately gives less efficient parameter estimates ($\hat{\boldsymbol{\beta}}$) and inefficient variance estimates. Liang and Zeger (1986) extended the quasi-likelihood approach to multivariate data by assuming that the marginal distributions are of the exponential family type and derived generalised estimation equations (GEE or GEE1). As with quasi-likelihood, the link function and the mean-variance relationship need to be specified, but also the correlation structure, which is assumed to depend on a parameter vector $\boldsymbol{\alpha}$. For some distributions, such as the multinomial distribution, the likelihood functions of a multivariate GLM (MGLM) (Fahrmeir and Tutz 2001) and the GEE are identical, but only if the correlation structure is correctly specified. The true correlation is usually unknown and it requires a "working guess" for the structure. Common *work-*

ing correlation structures are *unstructured*, *exchangeable*, *independent*, etc. The *naive* variance naively treats the working correlation as the true correlation structure. If the working correlation structure is indeed correct, the naive variance gives good estimates, otherwise it performs poorly, and instead a *robust* or *sandwich* variance is proposed (Liang and Zeger 1986). Independently of the choice of the correlation structure, the parameter estimates $\hat{\boldsymbol{\beta}}$ are consistent, given the model for the mean responses is correct. Under independence of items and an independence working correlation, the GEE are identical to the likelihood equations of a GLM (and the quasi-likelihood equations). For GEE, the observed vector $\mathbf{y}_i$ is often referred to as the ith *cluster*, with the components of the observed vector referred to as *observations*. For the remainder of the thesis $\mathbf{y}_i$ will be referred to as the ith observation or ith multiple response.

For binary response vectors, Prentice (1988) extended the GEE method to simultaneous modelling of the mean responses and modelling of the correlations. The correlation model parameters $\boldsymbol{\alpha}$ are obtained from a second set of estimating equations, which is of the same form as the first set of estimating equations with model parameters $\boldsymbol{\beta}$. Only the ith residual vector is replaced by the vector of differences between the empirical and true pairwise correlations. The obtained parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ are orthogonal. If the correlation model is wrongly specified, $\hat{\boldsymbol{\alpha}}$ is not consistent anymore, however, $\hat{\boldsymbol{\beta}}$ is still consistent provided the model for the mean responses is correct. Zhao and Prentice (1990) also modelled the mean response and correlation parameters for correlated binary data assuming the joint distribution is a member of the quadratic exponential family. The density of a member of this familiy has the form: $f(\mathbf{y}_i) = \Delta_i^{-1} \exp(\mathbf{y}_i^T \boldsymbol{\theta}_i + \boldsymbol{\omega}_i^T \boldsymbol{\lambda}_i + c_k(\mathbf{y}_i))$, where $\boldsymbol{\theta}_i$, $\boldsymbol{\omega}_i$ and $\boldsymbol{\lambda}_i$ are vectors of canonical parameters, $\Delta_i = \Delta_i(\boldsymbol{\theta}_i, \boldsymbol{\omega}_i)$ is a normalising constant and $c_k(\cdot)$ is the

shape function. As Prentice (1988), they derived two sets of estimating equations, but by treating both the mean response and correlation model as one. Therefore, their two sets of estimating equations are different from those presented by Prentice, since they do not treat empirical correlations and observations as independent. The parameter estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ are not orthogonal anymore. This method only provides consistent estimates if both models involving $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are correctly specified. The approach of Zhao and Prentice (1990) is referred to as GEE2, whereas the approach by Prentice (1988) and Liang and Zeger (1986) is referred to as GEE1. If both models are correct, the estimates of GEE2 are more efficient than those of GEE1. Prentice and Zhao (1991) extended GEE2 to a wider class of distributions other than the correlated binary distribution. Obtaining efficient GEE2 estimates depends on the correct specification of the third and fourth order central moments of $\mathbf{y}_i$ which is analogous to the correct specification of the working correlation (respectively of the second order moments of $\mathbf{y}_i$ for GEE1).

Lipsitz et al. (1991) and Liang et al. (1992) use the odds ratio instead of the correlation coefficient as a measure of association. Fitzmaurice and Laird (1993) derived likelihood equations assuming $\mathbf{y}_i$ is from the quadratic exponential family. They modelled the mean response, but used the conditional log odds ratios as the association parameter. Their iteration scheme uses a Fisher scoring algorithm, where, for each step, the iterative proportional fitting (IPF) algorithm (Bishop, Fienberg and Holland 1975) is applied to obtain updates of the higher order moments, which are required for the computation of the likelihood. The mean response model parameters are robust provided the mean response model is correctly specified independently of the association model. Heagerty and Zeger (1996) investigated mean response and association models for clustered ordinal responses. They considered the global odds ratio and the correlation coefficient as

possible association parameters, and derived GEE2 based on a general log-linear model representation of the likelihood of a single cluster by setting its higher order parameters to zero. Generally, in order to obtain the full likelihood, we need to specify all parameters up to the Jth order of the joint distribution, but the mean response (and association) model only provides us with first (and second) order parameters. Fitzmaurice and Laird (1993) circumvented this difficulty of computing the higher order parameters directly by using the IPF algorithm and computing these parameters indirectly, which involves a high computational burden for each step. The benefit of their method is that it yields real ML estimates, in contrast to the other GEE methods although some of those were partly derived from the likelihood equations for the quadratic exponential family. Carey, Zeger and Diggle (1993) used another approach called alternate logistic regression (ALR) also using the odds ratio as a measure of association. Their method uses the same estimating equations for parameters $\boldsymbol{\beta}$ as GEE1. However, the odds ratio arises in the conditional expectation using a unique approach of unbiased nonlinear estimating equations. The authors report high efficiency of both mean and association model parameters while retaining robustness of $\hat{\boldsymbol{\beta}}$. Heagerty and Zeger (1996) extended this method for ordinal responses yielding slight efficiency advantages in the estimation of $\boldsymbol{\alpha}$ over GEE1.

Let now index k stand for the kth group or covariate setting. For instance, Table 1.1 has $K \times r = 2 \times 2 = 4$ covariate settings. We can also express the marginal model in terms of the expected joint (table) counts $\mathbf{m}_k$ and function $\mathbf{L}$ by $\mathbf{L}(\mathbf{m}_k) = \mathbf{Z}_k\boldsymbol{\beta}$, because $\boldsymbol{\mu}_k$ can be easily computed from $\mathbf{m}_k$. ML estimation for generalised log-linear models (GLLM) of the form $\mathbf{L}(\mathbf{m}_k) = \mathbf{C}\log\mathbf{M}\mathbf{m}_k = \mathbf{Z}_k\boldsymbol{\beta}$ were considered by Lang and Agresti (1994) and Lang (1996) by applying a constrained equation specification of the model for which Aitchison and Sil-

vey (1958, 1960), Silvey (1959) and Aitchison (1962) laid out much of the theoretical foundation, where the model is re-expressed as a system of constraints and Lagrange multipliers. The common freedom specification of the model as for GEE or standard ML estimation does not allow the joint parameter to be expressed in terms of the modelled parameter, because there is a many-to-one relationship. Haber (1985) considered ML estimation for linear link and a special class of GLLM. Log-linear, logit, cumulative logit and multivariate logit models (McCullagh and Nelder 1989, Glonek and McCullagh 1995, Glonek 1996) are subclasses of GLLM. Lang (2004) extended this class to multinomial-Poisson homogeneous (MPH) models by outlining a general theory of the constraint approach for contingency table models. Lang (2005) introduced homogeneous linear predictor (HLP) models, an important subclass of MPH models, which have the form $\mathbf{L}(\mathbf{m}_k) = \mathbf{Z}_k\boldsymbol{\beta}$. Linear predictor models (Bergsma 1997) are formally equivalent to HLP models, however they are implicitly restricted to allow asymptotic approximations in contrast to HLP models. For our type of marginal modelling, GLLM only allows the logistic link, in contrast to HLP, which also allows other (smooth) popular links such as the probit link.

1.1.3 Random Effect Approach

The marginal model approach applies directly to the marginal distributions of $\mathbf{y}_i$. The parameters $\boldsymbol{\beta}$ in a GLM or GEE are called *fixed effects* and are independent of the sample. This type of modelling is called *population-averaged*. In contrast, generalised linear mixed models (GLMM) additionally include a *cluster specific effect* or the *random effect*. Conditional on the random effect $\mathbf{u}_i$, the distribution of $\mathbf{y}_i$ is assumed to be of the exponential family and has the form $\mathbf{g}(\boldsymbol{\mu}_i|\mathbf{u}_i) = \mathbf{Z}_i\boldsymbol{\beta} + \mathbf{Q}_i\mathbf{u}_i$, where $\mathbf{Q}_i$ denotes the contribution of the ith subject to a design matrix $\mathbf{Q}$ for the

random effects. The distribution of $\mathbf{u}_i$ is often assumed to be multivariate normal with zero mean and variance $\mathbf{\Sigma}$. This model approach is also called *cluster-* or *subject specific* since it accounts for subject specific mean responses. For normally distributed data, the corresponding linear mixed models have been extensively developed after some seminal papers (Harville 1976, Laird and Ware 1982). Obtaining parameter estimates for a linear mixed model for the fixed effect parameters, their variance and the random effect variance is relatively simple and the estimates of the fixed parameters have even closed forms.

An early application of a GLMM is the *Rasch model* (Rasch 1961), modelling binary correlated data by a simple logistic random effect model, where the estimates are obtained through conditional ML. GLMMs were also used to account for over-dispersion in binomial (Williams 1982) and Poisson (Breslow 1984) regression models. Agresti et al. (2000) describe various social science applications of GLMM. Agresti and Natarajan (2001) review developments in random effect models for ordinal data, whereas Hartzel, Agresti and Caffo (2001) discuss GLMM methods for nominal outcomes. ML estimation is accomplished by integrating over the random effect distribution. As a result, ML estimation is much more complicated. The most frequently used methods are based on first- and second order Taylor series expansions. Marginal Quasi-likelihood (MQL) involves expansion around the fixed part of the model, whereas penalised quasi-likelihood (PQL) also includes the random part in its expansion. Stiratelli, Laird and Ware (1984) derive an approximate Bayes procedure which is identical to a PQL approach suggested by Schall (1991) and Breslow and Clayton (1993). Several authors (Zeger, Liang and Albert 1988, Goldstein 1991) used MQL to focus on the marginal relationship between covariates and outcome. Unfortunately, PQL and MQL methods yield estimates that are biased towards zero in several situations,

in particular for first order expansions (Breslow and Lin 1995). Raudenbush, Yang and Yosef (2000) introduced a fast method combining a fully multivariate Taylor series expansion and a Laplace approximation, yielding accurate results. Also, in contrast to PQL and MQL, the deviance obtained from their method can be used for likelihood ratio tests.

Another method to obtain real ML estimates is numerical integration. If the random effect distribution is normal, any practical degree of accuracy of the integral can be obtained with Gauss-Hermite quadrature approximation by increasing the number of quadrature points. However, this number increases exponentially with the dimension of the random effect vector $\mathbf{u}_i$. Liu and Pierce (1994) and Rabe-Hesketh, Skrondal and Pickles (2002) considered adaptive Gauss-Hermite quadrature methods to reduce the number of quadrature points.

Due to a breakthrough in recent computer technology, iterative simulations can also be used to approximate the integral. Monte Carlo (MC) techniques are one useful tool to sample from the random effect distribution. If sampling from the random effect distribution is difficult, importance sampling as suggested by Geyer and Thompson (1992) and Gelfand and Carlin (1993) is an alternative method, which was termed simulated ML (SML) by McCulloch (1997). A very popular method is maximising the likelihood via the EM algorithm (Dempster, Laird and Rubin 1977). The EM algorithm consists of two steps: the E(xpectation)-step and the M(aximisation)-step. Both steps can be performed separately for the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, because the EM algorithm is based on the complete log-likelihood which can be decomposed in a sum of two terms, where the first term depends on $\boldsymbol{\beta}$ and the second on $\boldsymbol{\Sigma}$. Generally, the EM algorithm also requires to solve integrals numerically. The integrals are with respect to the conditional distribution of $\mathbf{u}_i$ given $\mathbf{y}_i$, which can be achieved by several MC Markov Chain

(MCMC) methods. McCulloch (1997) suggested one method called MCEM based on the Metropolis-Hastings algorithm, whereas Booth and Hobert (1999) considered the use of rejection sampling to yield real independent samples. McCulloch (1997) also proposed a MC Newton Raphson (MCNR) algorithm. Both MCEM and MCNR reach the neighbourhood of the ML estimates (MLE) quickly, however achieving high accuracy requires a rapidly increasing amount of time. In contrast, SML performs poorly when using the true distribution; an unknown optimal importance sampling distribution must be used to yield good estimates. Instead, McCulloch (1997) suggested a hybrid method starting with MCEM or MCNR to get rough estimates and then finishing with SML. The estimates of MCNR or MCEM can be used to approximate the optimal importance sampling distribution of SML. Another advantage is that the hybrid method yields an estimate of the likelihood as a by product, which is not available from MCNR or MCEM.

Booth and Hobert (1999) also suggested to construct confidence intervals of the estimates for each iteration of the EM algorithm to limit the number of points needed to approximate the integrals. Their algorithm additionally provides the information matrix using the formulae presented by Louis (1982). Tutz and Hennevogl (1996) considered random effects models for ordinal data by applying several EM algorithms. Their ML approach uses a parameter transformation, such that the components of the random effects vector are independent and normally distributed, to make the iterative procedure easier.

The random effect estimates cannot be obtained by using a frequentist approach, because it requires the knowledge of the conditional distribution of $\mathbf{u}_i$ given $\mathbf{y}_i$, which is unknown. However this conditional or posterior distribution can be obtained from a Bayesian approach by applying Bayes' theorem. Then an

estimate of the random effect $\mathbf{u}_i$ is the mean of a large sample of the posterior distribution of $\mathbf{u}_i$ given $\mathbf{y}_i$ for known parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. These parameters are unknown, but can be replaced by their estimates.

Generally, the subject-specific effects tend to be larger in absolute value than the population-averaged effects, but so do the standard errors (Agresti and Liu 2001), hence, messages regarding significance are similar. If the variance of the random effects is zero, the random effect model and the marginal model are identical. For nonzero variances, the implied marginal model for μ_{ij} does not have the same form as the random effect model. For instance, a logit random effect model does not imply a logit model for the marginal mean. However, Zeger et al. (1988) show the marginal mean of a logit random effect model can be approximated by a logit model.

Pure Bayesian mixed models also have great popularity. The prior distributions of all parameters must be specified in advance. Parameter estimates are obtained by sampling from the posterior distribution (Fahrmeir and Tutz 2001, Ch. 7).

1.1.4 Loglinear Models

The full likelihood of each cluster can be represented by a log-linear model specifying all $2^J - 1$ parameters, for instance, Heagerty and Zeger (1996) derived GEE2 for clustered ordinal data starting from a log-linear representation of the log-likelihood. Liang et al. (1992) compare the marginal and log-linear model approaches. The parameters of a log-linear model are interpreted in terms of conditional probabilities and the parameters of a marginal model refer directly to the marginal probabilities, which are expressed in terms of some explanatory variables. Agresti and Liu (2001) discussed several log-linear models for multiple

response data, such as independence models, assuming independence between items or assuming conditional independence of items given an explanatory variable. They also show a connection between the quasi-symmetry log-linear model and a simple random effect model. Despite log-linear models being fitted efficiently using the IPF algorithm (Bishop et al. 1975), the interpretation of their parameters is difficult and we cannot model the mean responses directly in terms of the values of the covariates, in contrast to the marginal model approach. The log-linear model approach seems sensible only for quite simple models, but for more complex questions, log-linear models are impractical. The next section reviews Mantel-Haenszel methods, which can also be seen as a modelling approach.

1.2 Review: Mantel-Haenszel (MH) Methods

In this section, we review Mantel-Haenszel (MH) methods, starting with ordinary MH methods for a series of 2×2 tables. Then we follow with MH methods for a series of $r \times J$ tables and finally discuss MH methods for multiple response data.

1.2.1 The Ordinary Mantel-Haenszel Method

The odds of an event (or condition) is defined by $\pi/(1-\pi)$, where π is the probability of the event. The odds ratio Ψ is the ratio of the odds of an event occurring in one group to the odds of that event in another group. These groups might be men and women, an experimental group and a control group, or any other dichotomous classification. The odds ratio is used to test whether the probability of a certain event is the same for two groups. We note that the odds ratio takes values in $(0, \infty)$. An odds ratio of 1 indicates that the event under study is equally likely in both groups. If $\Psi > 1$, then the event is more likely in the first

group, whereas $\Psi < 1$ indicates that it is less likely. The 2×2 Table 1.2 shows observations for two such groups and events A and $\bar{A}$, the complement of A.

Table 1.2: 2×2 Table for Event A and 2 Groups

	A	$\bar{A}$	totals
group 1	X_1	$n_1 - X_1$	n_1
group 2	X_2	$n_2 - X_2$	n_2
totals	$X_1 + X_2$	$n_1 + n_2 - X_1 - X_2$	$n_1 + n_2$

The odds ratio $\Psi = \pi_1(1-\pi_2)/\{\pi_2(1-\pi_1)\}$ is estimated by $X_1(n_2 - X_2)/\{X_2(n_1 - X_1)\}$, which is invariant if rows or columns (or both simultaneously) are interchanged. In clinical studies there are often only a few subjects. Multicentre trials increase the sample size, but populations differ for different centres and one cannot assume that probabilities for different centres are equal. However, one can assume that the odds ratios for each of the K centres are identical, that is, assuming a *common odds ratio* Ψ with $\Psi = \Psi_1 = \cdots = \Psi_K$. Under this *common odds ratio assumption*, the *Mantel-Haenszel (1959)* estimator $\hat{\Psi}$ of the common odds ratio is widely used by practising statisticians and epidemiologists. The MH estimator is a ratio of two sums C_{12} and C_{21}, where each summand of C_{ij} has the form $X_{ik}(n_{jk} - X_{jk})/(n_{ik} + n_{jk})$ with index k referring to the quantities of the kth table or kth centre. The factor $1/(n_{ik} + n_{jk})$ is a weight accounting for the sample size of the kth table. The MH estimator is also often applied for other stratified data for which the common odds ratio assumption is reasonable.

Even if the assumption of a common odds ratio is slightly violated, the MH estimator is still a useful tool to summarise the association across tables. Despite the Mantel-Haenszel estimator's simplicity, it has some useful properties. First, it applies to very sparse data. More precisely, it is defined when only one summand

of C_{12} and of C_{21} is non-zero.

It is also *dually consistent*, that is, consistent under two types of asymptotic models: (1) when the sample size of each stratum increases and the number of strata is fixed, and (2) when the number of observations becomes large with the number of strata, while the sample size of each stratum remains fixed. We refer to (1) as a *large-stratum* limiting model, or model I, and to (2) as a *sparse data* limiting model, or model II. In practice, model I represents large $n_{1k} + n_{2k}$ for each stratum and model II respresents large K. The MH estimator is robust under any such extreme data. The consistency of the MH estimator for model I was shown by Gart (1962) and for model II by Breslow (1981). Hauck (1979) derived the limiting variance of the MH estimator under model I, whereas Breslow (1981) derived two asymptotic variances under model II: one based on the conditional distribution of the observations for each table given the marginal totals, and the other on the empirical variance. Applying either of the variance estimators depending on the given data, whether the data resembles the sparse data or large stratum case, is very unsatisfactory. Breslow and Liang (1982) proposed a weighted average of the two variance estimators to account for the two different limiting models. Robins, Breslow and Greenland (1986) proposed a variance estimator which is dually consistent under models I and II based on the unconditional distribution of the data.

An alternative way to estimate the common odds ratio for K 2×2 tables is to fit an ordinary logit model with main effects and no interaction, where the K strata and one binary classification are treated as factors and the other binary classification as a response. The corresponding loglinear model is a model with no three-way interaction among rows, columns and strata. However, the unconditional maximum likelihood (ML) estimator is a poor estimator, because under

model II the nuisance parameters grow as the sample size grows. For instance when each table consists of a single matched pair, then the unconditional ML estimator of the common odds ratio converges to the square of the true common odds ratio (Anderson 1980, p.244). The nuisance parameters can be eliminated by conditioning on the margins of the 2×2 contingency table. The ML estimator based on the conditional distribution, which is noncentral hypergeometric in each stratum, is also dually consistent. As a by-product, the ML fitting yields a variance estimator of the odds ratio estimator.

Cochran (1954) introduced a statistic for testing conditional independence, that is independence of the variables forming the rows and columns of the tables, conditional on the K levels of the third variable. The test statistic is based on a weighted sum of table-specific differences in proportions conditioning on the row totals, supposing each 2×2 table consists of independent binomials. Mantel and Haenszel (1959) proposed a similar (MH) test statistic based on the hypergeometric distribution as the conditional ML estimator. These two statistics typically differ by a negligible term, and both are asymptotically chi-squared with 1 degree of freedom ($\chi^2(1)$). They are also known as Cochran-Mantel-Haenszel (CMH) statistics. The tests are inappropriate when the association changes significantly across strata. The MH estimator equals unity only if the MH test statistic equals zero. Hence a significance test using the MH test statistic can detect any departure from unity of the weighted average of the stratum-specific odds ratios.

If the assumption of a common odds ratio fails, we can still use the MH estimate as a summary of the odds ratios among the strata. Without the common odds ratio assumption, the MH estimator is consistent under model I only; and appropriate standard errors were suggested by Guilbaud (1983), since the dually consistent variance estimator of Robins et al. (1986) fails.

A simple way to test the homogeneity of the odds ratio across strata is to apply a goodness-of-fit test to a logit model with only main effects and no interaction. The goodness-of-fit test statistic has $K-1$ degrees of freedom (df) if the model holds. Breslow and Day (1980) developed a test statistic which does not require model fitting and focuses directly on the potential lack of homogeneity. The Breslow-Day test statistic sums the squared deviations of observed and fitted values each standardised by its variance. According to Breslow and Day (1980) the test statistic should follow a chi-squared distribution with $df = K-1$. Tarone (1985) proved that it is stochastically larger under the homogeneity assumption, and developed a modified score test statistic that is indeed asymptotically $\chi^2(K-1)$. A drawback of these methods is that they are inappropriate under model II. Instead, Liang and Self (1985) proposed a score test assuming the log odds ratios across strata are independent and identically distributed, which is valid also when the sample size increases with the number of strata. Paul and Donner (1989) conducted a simulation study generally recommending Tarone's modified test statistic. Liu and Pierce (1993) used a different approach by assuming that the log odds ratios across the strata are a sample from a population with unknown mean and variance. They investigated the conditional likelihood functions for the mean and the variance. A test of homogeneity of the odds ratios can be conducted by testing whether the variance of the log odds ratio equals zero. Liu and Pierce (1993)'s approach is more general than that of Liang and Self (1985), since it describes the heterogeneity of the log odds ratios across the strata.

1.2.2 Extended Mantel Haenszel Methods

The previous section discussed the MH method for binary responses only. Now we review methods for the multiple response case, generally forming one $r \times J$

table for each stratum. First we review the generalised MH estimator for K $2 \times J$ contingency tables with J nominal response categories. For columns x and y, one can obtain a partial MH estimator by applying the ordinary MH estimator for those two columns only. Mickey and Elashoff (1985) proposed a more efficient generalised partial log odds ratio estimator by using information from all pairs of columns. They introduced this generalisation to estimate the log odds ratio for a log-linear model with no three factor interaction, but their method is generally applicable to any partial log odds ratio estimator. Greenland (1989) extended their method to the log MH estimator from all 2×2 subtables per stratum, yielding the generalised MH estimator; and derived corresponding dually consistent variance and covariance formulas. Liang (1987) introduced a class of estimating functions by extending the Mickey and Elasthoff method, where the ordinary MH estimator is a special case. Sato (1991) derived dually consistent (co-)variance estimators from Liang's estimating functions approach. Yanagawa and Fujii (1995) proposed a projection method for K $2 \times J$ contingency tables. The method is applied to some arbitrary log odds ratio estimator to obtain an invariant log odds ratio estimator. For example, the projection method applied to the ordinary MH estimator yields the generalised MH estimator; and similarly using the log-linear model approach one obtains the generalised estimator proposed by Mickey and Elashoff (1985). Both Mickey-Elasthoff's and Greenland's generalised estimators are asymptotically equivalent to Liang's estimation functions with appropriate weighting. As in the binary case, conditioning on the marginal totals for each $2 \times J$ table, the counts follow a noncentral multiple hypergeometric distribution (Plackett 1981, p.81). The corresponding conditional ML estimators are also dually consistent, but they might impose a high computational burden (Mickey and Elashoff 1985).

For testing conditional independence, Birch (1965), Landis, Heyman and Koch (1978) and Mantel (1978) extended the MH test statistic to K $r \times J$ tables. Conditional on the margins of each table, the cells follow a multiple hypergeometric distribution under conditional independence. Given nominal responses the test statistic is chi-squared with $df = (r-1)(J-1)$ under both limiting models. For ordinal responses we can assign scores to the response categories and the column categories, and then the test statistic (Mantel 1963) is asymptotically $\chi^2(1)$. For ordinary multinomial responses, Zhang and Boos (1996) investigated several generalised (Cochran-)Mantel-Haenszel statistics testing independence between the treatment variable (r treatments) and the multinomial response variable (J categories)

Yanagawa and Fujii (1990) extended the Breslow and Day (1980) test for homogeneity of K 2×2 tables, to test the homogeneity of the partial odds ratios of K $2 \times J$ contingency tables. Following Tarone (1985)'s approach, they adjusted the Breslow-Day statistic to have asymptotically a chi-squared distribution with $df = (K-1)(J-1)$ if the common odds ratio assumption holds.

For ordinal K $2 \times J$ contingency tables the common ordinal odds ratio can be estimated by fitting a proportional odds model or an adjacent-category logit model with the ML approach, and assuming no interaction between row and stratum variables. However, ML estimation may not yield a good estimator if the data are sparse. Since the proportional odds model is not a canonical link model, the conditioning on the marginal tables does not eliminate the nuisance parameter. Hence there is no conditional ML estimation for the proportional odds model.

McCullagh and Nelder (1989, p.273) introduced a pseudo "conditional likelihood" estimate for the ordinal odds ratio in the one stratum case, based on

the proportional odds model and using an estimating equation. However, their method cannot be used for extremely sparse data, because it computes the inverse of the cell counts.

Hartzel, Liu and Agresti (2001) discuss several random effect models for ordinal data, such as the proportional odds model, the adjacent categories logit model and the loglinear model of heterogeneous linear-by-linear association. The fixed effects summarise the actual effect, while simultaneously the random effects describe the degree of heterogeneity across strata.

Clayton (1974) provided a more complex estimator of the log common ordinal odds ratio, based on a weighted average of estimators and a separate collapsing of each partial table. However, it remains unclear how to construct sparse data standard errors. The MH estimator was generalised by Liu and Agresti (1996) for K ordinal $2 \times J$ tables. They derived a dually consistent ordinal common odds ratio estimator and also a dually consistent variance estimator. This ordinal common odds ratio estimator simplifies dramatically for matched pairs to an estimator, which was previously proposed by Agresti and Lang (1993).

Liu (2003) extended the MH type common ordinal odds ratio estimator to K $r \times J$ contingency tables, where J is the number of ordered response categories and r the number of categories of an explanatory variable. Liu also provided not only dually consistent (co-)variance estimators, but also generalised estimators following the Mickey and Elashoff (1985) approach. For K $r \times J$ tables, the common ordinal odds ratio is also known as the local-global odds ratio. It is local in the explanatory variable, because we can compare any two levels of the row variable. On the other hand it is global since it is based on all dichotomous collapsings of the response (J levels) for which each collapsed response has a binary outcome ($\leq j$,$> j$). When both the response and explanatory variables are ordi-

nal, the global odds ratio can be considered. Liu (2003) considers a model with a constant odds ratio for all strata and dichotomous collapsings of the responses into a pair of binary outcomes ($\leq j$,$> j$) and ($\leq i$,$> i$). The odds ratio is referred to as a *global odds ratio*, because it describes the conditional association between the two variables globally. One way to obtain a global odds ratio estimate is to use ML estimation for homogenous linear predictor models (Lang 2005). Liu (2003) also proposes another dually consistent MH type estimator for the global odds ratio. Since the dually consistent variance estimator is too complex to derive, Liu proposed a bootstrap estimate of standard error. Liu and Agresti (1996) and Liu (2003) also introduce a Wald statistic to test the homogeneity across strata of the local-global and global odds ratios.

Liu (1995) introduced a test of conditional independence under the assumption of a common cumulative odds ratio. Liu also follows the approach by Liu and Pierce (1993) assuming that the cumulative odds ratio behaves like a sample from a population with unknown mean and variance.

Liu and Wang (2007) considered two diagnostic strategies for evaluating the heterogeneity of the ordinal odds ratios across strata. The first strategy uses a proportional odds model allowing random effects, where the standard deviations of the random effects measure the heterogeneity of the ordinal odds ratios. The second approach uses the Cook (1977) distance applied to the MH type estimator of the ordinal odds ratio as a measure of influence. It shows in detail the heterogeneity of each stratum.

1.2.3 Extending the Mantel-Haenszel Method to Multiple Response Data

Many surveys allow each respondent to tick any number out of J categories. This multiple outcome variable is referred to as pick any/J variables and the corresponding data as pick any/J data, where "/" stands for "out of" (Coombs 1964). Each of the J category responses is called an item (Agresti and Liu 1998). A problem of interest is whether responses to each category are marginally independent of the row variable with r levels in the absence of a stratification variable, i.e. a $r \times J$ table, where the rows refer to a group variable and where the columns refer to the items of a multiple response variable. Agresti and Liu (1999) called this *multiple marginal independence* (MMI). Bilder, Loughin and Nettleton (2000) reviewed several existing methods for testing MMI and conducted a simulation study investigating their performance. They found that the best results came from a naive sum statistic proposed by Agresti and Liu (1999), which is a symmetric version of a test originally proposed by Loughin and Scherer (1998). Since the distribution of the test statistic is unknown, they used bootstrapping and a newly proposed p-value combination method to obtain its distribution.

Bilder and Loughin (2002) investigated MMI in the presence of a stratification variable with K levels, that is conditional MMI (CMMI). They proposed an extended Cochran statistic which is chi-squared with $df = (J-1)(r-1)$ when the items are independent and suggested bootstrapping to obtain the distribution of the test statistic when the items are dependent. Bilder and Loughin (2004) investigated marginal independence between two categorical variables and proposed a modified Pearson statistic following the approach by Loughin and Scherer (1998). Again, bootsrapping is suggested to obtain the sampling distribution of the test

statistic.

1.3 Review: Diagnostics Methods

Regression models are characterised by a relationship between the covariates $\mathbf{x}_i$ and the J dimensional observations $\mathbf{y}_i$ $(i = 1, \ldots, n)$. For example a GLM or a marginal model for multiple response data can be expressed as $\mathrm{g}(\boldsymbol{\mu}_i) = \mathbf{Z}_i\boldsymbol{\beta}$ with previously introduced notations, where $\boldsymbol{\beta}$ is a p-dimensional parameter vector. The model assumes that all n observations follow the same law or distribution, i.e. $\mathrm{g}(\boldsymbol{\mu}_i) = \mathbf{Z}_i\boldsymbol{\beta};\ \ i = 1, \ldots, n$. However, in reality, this is a very restrictive and unrealistic assumption, because it seems likely that some observations do not follow the model, which might lead to wrong statistical inference. Other observations might follow the model, but may possibly falsify results due to, for example, the extremeness of the sampled values. The main goal of diagnostic methods is to detect such observations and to eliminate them, to avoid seriously misleading representation of the data.

1.3.1 Linear Models

A linear model has the general form $\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ are the error terms, most commonly with independent zero mean random variables and common variance σ^2. Quantities such as vector $\mathbf{y}$ and design matrix $\mathbf{Z}$ without index i refer to the stacked version containing all observations, e.g. $\mathbf{Z}$ stands for the design matrix $(\mathbf{Z}_1^T, \ldots, \mathbf{Z}_n^T)^T$. Also note that we sometimes use multivariate quantities such as $\mathbf{y}_i$ although observations are univariate in some instances. Cook and Weisberg (1982) comprehensively reviewed regression diagnostics for linear models with univariate observations ($J = 1$), which were well established by then. Some of

these statistics are briefly introduced. The residual vector $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$ describe the deviations of the observed data from the fit. Points with large residuals, representing model failure, are called *outliers*. The *leverage*, *hat* or *prediction matrix* maps $\mathbf{y}$ into $\hat{\mathbf{y}} = \boldsymbol{\mu}$ by $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ with $\mathbf{H} = \mathbf{Z}(\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{V}^{-1}$, where $\mathbf{V}$ is the covariance matrix of error terms $\boldsymbol{\epsilon}$. For independent and univariate observations $\mathbf{V} = \sigma^2\mathbf{I}$, for multivariate independent observations $\mathbf{V}$ is block-diagonal, and for the general linear model $\mathbf{V}$ has an arbitrary structure. Matrix $\mathbf{H}$ can also be seen as a *projection* matrix, because it generates the perpendicular projection of $\mathbf{y}$ into a $p \times J$-dimensional subspace. The leverage of the ith observation vector is defined as the trace of the corresponding submatrix $\mathbf{H}_i$ of $\mathbf{H}$. The leverage for the jth observation within $\mathbf{y}_i$ is the jth diagonal element of $\mathbf{H}_i$ or the simply the corresponding diagonal element of $\mathbf{H}$.

Leverage points are observations with a high leverage usually using $2p/n$ as a calibration point (Hoaglin and Welsch 1978). High leverage points can also be thought of as outliers with respect to the predictors, whereas outliers refer to model failure in the response variable. Preferable to the residuals are scaled residuals standardised by the variance or the leverage, as the *studentised* or *Pearson residuals*. Outliers occur frequently in real data, and they get often unnoticed, because nowadays data is usually processed by computers without further careful inspection or screening. Although the residuals and the leverage are effective in detecting extreme points, they cannot detect the impact of the extreme points on the estimates, residuals, etc. Other approaches are more useful: the deletion approach and the perturbation approach.

Observations whose inclusion or exclusion results in substantial changes for the fitted model are said to be *influential* or as Belsley, Kuh and Welsch (1980) formulate: "An influential observation is one which, either individually or to-

gether with several other observations, has demonstrably larger impact on the calculated values of various estimates ... than is the case for most of the other observations". Chatterjee and Hadi (1986) point out, that an "observation ... may not have the same impact on all regression outputs. The question 'Influence on what?' is, therefore, an important question." The observation might be influential on the parameter estimates $\hat{\boldsymbol{\beta}}$, on the residuals, or on the fitted values, etc. First, we point out, that neither an outlier nor a high leverage point needs to be influential. The influence function introduced by Hampel (1974) is a measure of influence on a statistic when adding a observation $(\mathbf{y}_i^T, \mathbf{x}_i^T)^T$ to the sample coming from a c.d.f. and computing a certain limit.

Let subscript $[i]$ denote the quantities with the ith observation being removed from the sample, e.g. $\hat{\boldsymbol{\beta}}_{[i]}$ denotes the parameter estimates from $n-1$ observations denoted by $\mathbf{y}_{[i]}$ excluding the ith observation $\mathbf{y}_i$. If i refers to a single observation, then the deletion is also called *single case deletion*, whereas if i is replaced by a set d we speak about *multiple case deletion*. The influence function on $\hat{\boldsymbol{\beta}}$ (or generally on a statistic T) with empirical c.d.f. denoted by F without computing the limit yields the *sample influence curve* or *function*, whereas the influence function on $\hat{\boldsymbol{\beta}}_{[i]}$ with empirical c.d.f. $F_{[i]}$ and computing a limit yields the *sensitivity curve* (Chatterjee and Hadi 1986). Both, the sample influence curve and the sensitivity curve are proportional to $\text{DBETA}_i := \Delta_i\hat{\boldsymbol{\beta}} := \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]}$. An important diagnostic measure is the Cook distance $CD_i := \Delta_i\hat{\boldsymbol{\beta}}^T \text{Cov}(\hat{\boldsymbol{\beta}})^{-1}\Delta_i\hat{\boldsymbol{\beta}}/p$, which is the Mahalanobis distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{[i]}$ with covariance matrix $\text{Cov}(\hat{\boldsymbol{\beta}})$ divided by p. The Cook distance was originally introduced to assess the influence on the confidence ellipsoid (Cook 1977). Following the above derivations from the influence curve, the Cook distance measures the influence of the ith observation on the parameter estimates $\hat{\boldsymbol{\beta}}$. For linear models, the Cook dis-

tance can also be re-expressed in terms of the leverage or in terms of the residuals as $(\Delta_i\hat{\mathbf{y}})^T(\Delta_i\hat{\mathbf{y}})/(p\hat{\sigma}^2)$. Neither high leverage points nor outliers need to be influential nor need to have large Cook distance values, but usually the larger the residuals and the leverage are, the larger the Cook distance is. The Cook distance does not follow exactly the F-distribution for a linear model, but generally, observations whose Cook distance is larger than two, should be carefully checked. There are also partial measures, such as the partial Cook distance or the partial leverage, investigating the effect on the jth parameter estimate or the effect of the jth covariate, etc. Chatterjee and Hadi (1986) review numerous influence measures for linear models, such as the Welsch-Kuh distance or the Welsch distance, and concluded that only a handful of methods are needed to assess influential observations. Case deletion methods for linear mixed models were considered by Christensen, Pearson and Johnson (1992).

Pena and Yohai (1995) introduced an influence matrix where the ijth entry has the form $(\Delta_i\hat{\mathbf{y}})^T(\Delta_j\hat{\mathbf{y}})/(p\hat{\sigma}^2)$. Clearly the diagonal elements are identical to the Cook distance. Their procedure aiming at detecting influential subsets is based on the analysis of the eigenstructure of the influence matrix. Lawrance (1995) used the conditional Cook distance $CD_{i[j]}$, that is deletion of case i after case j has been deleted, to investigate the effect of two observations. He compared $CD_{i[j]}$ with CD_i and CD_{ij} to distinguish interaction between a pair of cases. Interactions were categorised into five types, such as *swamping* and *masking*. Masking refers to a situation, when outliers are not detected due to multiple outliers interacting with each other, and swamping refers to the opposite, when the data wrongly suggests that a good point is an outlier. These are joint effects of observations that make the detection of influential observations difficult.

Munoz-Pichardo et al. (1995) proposed a different approach by studying the

influence in a general linear model with uncorrelated errors, based on the conditional bias. Banerjee and Frees (1997) considered influence diagnostics for linear longitudinal models. Outliers in linear multilevel models were considered by Langford and Lewis (1998).

Haslett (1999) focused on the conditional residuals for the class of general linear models with correlated errors. Special cases of this model are: The classic linear model with independent observations, longitudinal linear models, multivariate linear models, linear mixed models etc. The conditional residuals for a given set i of observations are the differences between the observations $\mathbf{y}_i$ and the best linear unbiased predictor of $\mathbf{y}_i$ given $\mathbf{y}_{[i]}$ provided a general estimate of the correlation matrix $\mathbf{R}$ is used, that is, the estimate of $\mathbf{R}$ is not refitted for each deletion. The substitution of the response $\mathbf{y}_i$ by its best unbiased predictor is also called "delete = replace" method. Multiple deletion diagnostics and deletion diagnostics based on the conditional residuals for the general linear model were also considered by Baade and Pettitt (2000). Haslett and Dillane (2004) proposed the same "delete=replace" method for linear mixed models focusing on deletion diagnostics for variance component estimation.

Haslett and Haslett (2007) gave a detailed review about three basic types of residuals for the general linear model: The marginal ($\mathbf{r}_i$), the conditional and the model specific residuals. They show a linear relationship between these three types of residuals, though they are essentially different. Generally, the "delete = replace" method provides a fast computational method to compute deletion diagnostics, which can be expressed as a linear function of the conditional residuals.

Zewotir and Galpin (2006) investigated the performance of deletion diagnostics for linear mixed model using a Monte-Carlo simulation study. Based on a

sensitivity analysis for the deletion diagnostics the authors obtain helpful results for the analysis of influential observations.

1.3.2 Generalised Linear Models and Extensions

Pregibon (1981) considered diagnostics for logistic regression for binary responses. In contrast to linear models, the fitting requires an iterative algorithm, such as Fisher scoring, and so does the computation of the various deletion diagnostics also require a refit of the model for each deleted observation. Cook and Weisberg (1982) and McCullagh and Nelder (1983) discussed deletion diagnostics for GLM, but Pregibon's one step approximations were not mentioned. Williams (1987) focused on deriving (one-step) approximations of deletion diagnostics for GLM. For a GLM, the difference $2\{l(\hat{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}_{[i]})\}/p$, where l denotes the likelihood, can be approximated by the Cook distance. Generally, the Cook distance gained acceptance as an influence measure for other models as well as GLM, because of its easy formulation in terms of DBETA$_i$, the covariance of parameter estimates and the number of parameters p, and its interpretation as a measure of influence on the parameter estimates, which is the primary interest of the practitioner.

Preisser and Qaqish (1996) investigated deletion diagnostics for generalised estimation equations, which is a broader class of models including univariate and multivariate GLM (Fahrmeir and Tutz 2001) as a subclass. The authors derived one-step formulae for DBETA and the Cook-distance for deleting an arbitrary set of observations and the sub-cases of deleting a cluster and an observation within a cluster. Ziegler et al. (1998) considered deletion diagnostics for the GEE1 approach for mean response models, but also for correlation models. However, they did not provide estimates for the correlation model parameters deleting the ith cluster. Such a formula was then provided by Preisser and Perin (2007).

Lee and Fung (1997) focused on the detection of multiple outliers in a GLM and non-linear regression. A stepwise procedure is proposed which might get control over commonly occurring problems like masking or swamping. Xiang, Tse and Lee (2002) investigated the Cook distance for GLMM deriving first order approximations for the best linear unbiased predictor, that is, they used a quasi-likelihood method to derive approximations, because methods based on real ML estimation are computationally expensive. Wang, Critchley and Liu (2004) considered diagnostics and influence analysis using a perturbation approach for the clustered sampling model for binary data.

1.3.3 Other Models and Methods

It seems the literature provides innumerable articles about diagnostic methods. We want to outline only few of them. Simonoff and Tsai (1991) investigated the influence on a goodness-of-fit test based on non-parametric regression. Fung et al. (2002) focused on influence diagnostics for semi-parametric mixed models based on maximum penalised likelihood estimation, extending the linear model framework. Deletion diagnostics for non-linear structural equation models were proposed by Lee and Lu (2003). They computed one step-approximation diagnostics based on the conditional expectation of the complete likelihood in the EM algorithm. Lee and Xu (2003) proposed a similar method deriving diagnostics for factor analysis models and ordinal categorical data.

Atkinson and Riani (1997) suggested a robust method to pinpoint influential observations for binomial data based on the forward search, which orders observations "from those most in agreement with the GLM to those least in agreement with." Their method is effective in pinpointing masked multiple outliers. Fay (2002) proposed a simple method to measure the effect of a single binary re-

sponse on logistic regression. After changing the binary response, the model is refitted and the change in the statistic of interest T is recorded, which is called *range of influence (ROI) for T*. The larger the ROI on T relative to the other responses is, the higher the influence of this changed binary response is. Wang, Jones and Storer (2006) compared two commonly used methods for case-deletion for the Cox-regression model: The empirical influence function approach and the covariate-vector approach, which outperforms the former according to their simulations.

1.3.4 Graphical Methods

By looking only at the numerical values of the deletion diagnostics, influential points can easily be undetected. Often, plots provide better insight and are an indispensable tool to detect such influential points. Chatterjee and Hadi (1988) considered a variety of common plots. For the simple linear regression model, the most effective technique for checking the assumption of the model is to make a scatterplot of a covariate versus the response and a residual plot of a covariate versus the residuals. Departures from linearity suggest that the model is not adequate. For generalised linear models, we would plot the inverse of the link function of the covariate versus the responses or the residuals. To visualise outliers and leverage points other plots are of importance: Regression diagnostic plot, a plot of the standardised residuals versus their index or versus their fitted values, a Normal Q-Q plot of the standardised residuals, a distance-distance plot, a leverage versus residual-squared plot, etc. or simply plotting one type of deletion diagnostics versus their index.

Lin and Wei (1991) proposed a lack-of-fit test for GLM, which is a normalised sum over those residuals for which a specific covariate is less or equal a certain

value t. The value t is arbitrary and the sum can then be thought of an empirical process or cumulative residual process, which converges to a Gaussian process with zero mean and some unknown covariance. There is another process which converges to the same limiting process and from which a sample can be obtained. Hence, the distribution of the first process can be approximated by a sample from the second. A supremum test, where large values indicate a lack-of-fit, can detect any departures from the functional form of the investigated covariate. In a similar manner, the overall model adequacy and the link function can be tested.

Lin, Wei and Ying (1993) used the same methods in a similar fashion to check the Cox model with cumulative sums of martingale-based residuals. Spiekerman and Lin (1996) extended the idea for the marginal Cox model for correlated failure time data and Lin and Spiekerman (1996) did similarly for parameteric regression from censored data. The supremum test yields small p-values if the residual process gives relatively large values. Another possibility to assess the significance is to plot the residual process versus the covariate along with a small sample of the second process that is asymptotically equivalent. If the curve of the observed residual process is relatively large in absolute value to the other curves of the simulated processes, then it indicates a violation of the model. The plotting of the processes might give more insight into a misfit of the model than the sheer p-value and serves as an excellent graphical diagnostic tool. Lin, Wei and Ying (2002) applied the cumulative residual processes to GEE to cover a wider range of models and Pan and Lin (2005) did similarly for GLMM. Another application of those cumulative residuals are stratified case control studies (Arbogast and Lin 2005) for which standard ML estimation for a logistic model yields biased intercept parameters but still allows valid inference for the regression coefficients.

There are also various other techniques introduced; we briefly outline now

two of those. Cook and Weisberg (1997) discussed graphical methods for nearly any kind of regression model. Their basic idea is to examine the fit of the model by using a series of marginal model plots, where on each of these plots nonparametric estimates from the model are compared with the estimates of a nonparametric fit. Pardoe and Cook (2002) assessed the adequacy of a logistic model by applying a graphical method based on the Bayesian framework.

In the next section, we focus on reviewing the proportional odds model and graphical diagnostic methods for detecting a wrong model specification of the proportional odds model.

1.4 Review: Proportional Odds Model

A response variable with more than two mutually exclusive categories is called a *polytomous variable*. Such a categorical variable is known as a *nominal variable* if the categories are not ordered, or as an *ordinal variable* if only the order matters but not the difference between its values. Examples of ordered categories are: Patient condition (good, fair, serious, critical), migraine severity or degree of pain (none, mild, moderate, severe), and playing ability for any sport (weak, average, strong, professional).

Let π_{ij} denote the probability that outcome category $j = 1, \ldots, J$ is observed for subject $i = 1, \ldots, n$. Then the jth cumulative probability is defined as $\pi^*_{ij} = \pi_{i1} + \pi_{i2} + \cdots + \pi_{ij}$. Most models for ordinal responses apply a link function such as the logit or probit link on the cumulative probabilities. Currently the most popular model for ordinal responses is the *proportional odds model*, which uses logits of the cumulative probabilities, also termed *cumulative logits*. This approach was first addressed by Williams and Grizzle (1972) and Simon (1974),

but did not gain much popularity until the seminal paper of McCullagh (1980). For the J-category ordinal variable Y and a corresponding set of predictors $\mathbf{x}$, a column vector, the proportional odds model has the form

$$\text{logit}[P(Y \leq j \mid \mathbf{x})] = \alpha_j - \mathbf{x}^T\boldsymbol{\gamma}, j = 1, ..., J-1,$$

with $\alpha_1 < \alpha_2 < \cdots < \alpha_{J-1}$. The parameters $\{\alpha_j\}$, also called *cut points*, are often of little interest and are usually regarded as nuisance parameters, in contrast to the parameter vector $\boldsymbol{\gamma}$. The minus sign in front of the predictor term allows the usual interpretation of each component of $\boldsymbol{\gamma}$, whether the effect is positive or negative, but this parametrisation is not necessarily needed. The model applies simultaneously to all cumulative probabilities assuming an identical effect on each cumulative probability. In particular, the odds ratio of cumulative probabilities $\text{logit}[P(Y \leq j \mid \mathbf{x}_{i_1})] - \text{logit}[P(Y \leq j \mid \mathbf{x}_{i_2})]$, also called the *cumulative odds ratio*, is identical for all responses j and any two subjects i_1 and i_2, and is proportional to the distance between $\mathbf{x}_{i_1}$ and $\mathbf{x}_{i_2}$. For more details of the cumulative odds ratio we refer to Sub-Section 1.2.2 on page 17. The proportional odds model received its name from this proportionality property which applies to each cumulative logit. If $J = 2$, the proportional odds model is simply the logistic regression model.

Assume that ordinal variable Y has an underlying continuous variable Y^* (Anderson and Philips 1981), which is called the *latent variable*. Also suppose the mean of Y^* is linearly related to $\mathbf{x}$ and that the variance of the conditional logistic distribution is constant. The cutpoints $\{\alpha_j\}$ provide intervals of the continuous scale. If the observations of Y^* are grouped according to these intervals, such that ordinal variable Y is obtained, then the effects of the proportional odds model are

proportional to those effects of the linear model involving Y^*.

There are many other cumulative link models which do not use a logit link but other smooth links, such as probit, log-log and complementary log-log. Another possibility is to apply the baseline-category logit model for ordinal data, which is usually used for nominal responses. Other ordinal models, such as the *continuation-ratio* logit model, were reviewed by Liu and Agresti (2005).

Generally, the proportional odds model and many other ordinal models can be written as $\mathrm{g}(\boldsymbol{\mu}_i) = \mathbf{Z}_i\boldsymbol{\gamma}$, which is the general form of a multivariate GLM (Fahrmeir and Tutz 2001), where $\boldsymbol{\mu}_i$ is the mean of the multivariate response $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})$ with $y_{ij} = 1$ if subject i selects category j and zero otherwise. Here we focus on the proportional odds model and leave other ordinal response models to the interested reader.

Peterson and Harrell (1990) and Brant (1990) fitted $J - 1$ separate logistic regression models for each dichotomisation of the response variable with effects $\boldsymbol{\gamma}_j$ and intercepts α_j. These $J-1$ logistic regression models are also referred to as the *partial proportional odds models*. Under the proportional odds model the null hypothesis is $\mathcal{H}_0 : \boldsymbol{\gamma}_1 = \cdots = \boldsymbol{\gamma}_{J-1}$. The authors propose tests such as the Wald test, the score test and the likelihood ratio test for the null hypothesis $\mathcal{H}_0$ to assess the validity of the proportional odds model. Stiger, Barnhart and Williamson (1999) proposed a Wald and score test to assess the proportional odds model assumption, applying the GEE methodology (Liang and Zeger 1986) . Their score test is based only on the proportional odds model, whereas the Wald test applies to the the fit of the partial proportional odds model.

Clustered polytomous data models can also be expressed in vector form as $\mathrm{g}(\boldsymbol{\mu}) = \mathbf{Z}\boldsymbol{\beta}$, $\mathrm{g}(\boldsymbol{\mu})$ standing for $(\mathrm{g}(\boldsymbol{\mu}_1)^T, \ldots, \mathrm{g}(\boldsymbol{\mu}_n^T))^T$ and $\mathbf{Z}$ for $(\mathbf{Z}_1^T, \ldots, \mathbf{Z}_n^T)^T$, and many of the aforementioned methods for fitting marginal models for multiple

responses, such as GEE, also apply here, see Section 1.1.2. Miller, Davis and Landis (1993) and Lipsitz, Kim and Zhao (1994) extended the GEE (also known as GEE1) approach for modelling correlated nominal and ordinal categorical data. Heagerty and Zeger (1996) applied a GEE2 approach (Zhao and Prentice 1990) for clustered ordinal data, which also includes the modelling of the association parameters.

Agresti and Lang (1993) considered a proportional odds model with random effects. Adding a random effect to a multivariate GLM defines a multivariate GLMM and the aforementioned methods (in Section 1.1.3 on page 8) are also applicable to random effect models for ordinal and clustered ordinal data. Tutz and Hennevogl (1996) and Hartzel, Liu and Agresti (2001) discussed several random effect models for clustered ordinal data and suggested several fitting procedures. Hartzel, Agresti and Caffo (2001) focused on random effect models for nominal and ordinal categorical variables. Random effect models for categorical data in the social sciences were reviewed extensively by Agresti et al. (2000).

Common methods to test the fit of a model compare observed frequencies with expected frequencies satisfying the model. Lipsitz, Fitzmaurice and Molenberghs (1996) generalised the Hosmer – Lemeshow statistic for testing the fit of a logistic regression model for binary data with continuous covariates to regression models for ordinal responses also with continuous covariates. They sum the components of the multivariate observations $\mathbf{y}_i$ and means $\boldsymbol{\mu}_i$ with some chosen coefficients λ_j (e.g. $\lambda_j = j$) to yield univariate quantities. The coefficients are called *scores* and the univariate mean obtained by the sum is now called the *mean score*. The data is partitioned into G regions according to that mean score. Now another ordinal model with the same cutpoints and the same effects is fitted, but also including coefficients that assign observations to their regions. Under the

null hypothesis that the original model is true, these coefficients are zero, regardless of how the regions or scores were chosen. Standard goodness-of-fit tests, such as the Wald-test, the likelihood-ratio test and the score test with $d.f. = G - 1$, are used to test whether the coefficients are zero. If this null hyposthesis is rejected, then so is the original model.

Toledano and Gatsonis (1996) generalised the receiver operating characteristic (ROC) curve that plots sensitivity against 1 - specificity for all possible collapsings of the J categories. Kim (2003) introduced a graphical method for assessing the proportional odds assumption, which plots the probabilities $P(Y = y_i|\mathbf{x}_i)$ of the proportional odds model versus $P(Y = y_i|\mathbf{x}_i)$ of the partial proportional odds model, where index i refers to the ith subject. If the proportional odds model holds, the points should lie on a line with slope 1. Then a *reference plot* is obtained by plotting the same probabilities versus each other, but the probabilities are obtained from an artificial sample from the proportional odds model with parameters equal to the ML estimates of the real data set. The points of the reference plot are expected to lie near the line with slope 1, since the artificial data set follows the proportional odds model assumption. The reference plot helps in evaluating whether the plots of the real data set indicates a violation of the model assumption.

Another important statistical issue was targeted by Perevozskaya, Rosenberger and Haines (2003). They investigated the D-optimal design for the proportional odds model.

McCullagh (1980) also considered the proportional hazard model, an important model in the analysis of survival data. Here the ratio of log-survivor functions depends only on the difference of two covariates, like the log odds ratio of cumulative probabilities for the proportional odds model. Bennett (1983b) pro-

posed another model, also called proportional odds model, for survival analysis, where the odds ratio $\theta(t)$ is a ratio of two cumulative distribution functions $F_i(t)$

$$\frac{F_i(t)}{1-F_i(t)} = \frac{F_j(t)}{1-F_j(t)}\theta(t),$$

where i and j might refer to two observations or groups. If the event is failure or death then $S_i(t) := 1 - F_i(t)$ is the survival function. Inference for this proportional odds model was considered by many authors Bennett (1983a), Pettitt (1984), Wu (1995), Murphy, Rossini and vanderVaart (1997), Yang and Prentice (1999), Kirmani and Gupta (2001), Hunter and Lange (2002), Zeng, Lin and Yin (2005), Sundaram (2006), Chen, Tong and Sun (2007), Sun, Sun and Zhu (2007), and Lu and Zhang (2007).

1.5 Outline of the Thesis

The first three chapters (Ch. 2, 3 and 4) investigate odds ratio estimation for K independent tables of multiple response data, where each table consists of r independent rows of multiple responses with J items each. Chapter 2 defines the odds ratio in terms of one item and two rows. The ordinary MH estimator and its variance estimator are still applicable owing to the independence of rows. However, two MH estimators referring to different items are not independent anymore. We derive new dually consistent estimators for the covariances between any two MH estimators. We also investigate the performance of the MH estimator under dependence between strata. Under this dependence, ML estimation is no longer as easy as under independence, but can be achieved by fitting a homogenous linear predictor (HLP) model (Lang 2005). Since for large J this method is infeasible, we use GEE (Liang and Zeger 1986) instead. We conduct a

simulation study to investigate the performance of the estimators.

In Chapters 3 and 4 we generalise to the multiple response case the two types of $K \times 2 \times J$ tables considered by Greenland (1989): (a) two rows of independent multinomials for each stratum, and (b) J independent binomials for each stratum. Chapter 3 extents case (a) to (a'): two rows of independent multiple responses with J categories per table, also forming a $2 \times J$ table. Then Chapter 4 extends case (b) to (b'): 1 row of multiple responses with J items (or equivalently J dependent binomials) forming a $2 \times J$ table. For case (b'), the ordinary MH estimator is no longer dually consistent, but consistent only under model I (Yanagawa and Fujii 1995). We propose a new dually consistent MH type estimator and also derive a dually consistent variance estimator for that new MH estimator. For case (a'), we prove that the ordinary MH estimator is dually consistent and derive new dually consistent (co-)variance estimators. We also propose a generalised estimator following the Mickey and Elashoff (1985) approach. For cases (a') and (b'), we conduct a simulation study confirming the dual consistency and good properties of the estimators. In addition, we consider a diagnostic strategy to detect heterogeneity of the estimators across the strata.

Chapter 5 investigates deletion diagnostics for multiple response data applying the GEE and HLP methodology. Preisser and Qaqish (1996) have considered deletion diagnostics for GEE, but deletion diagnostics for HLP models have not been considered previously. Methods are then illustrated using an example of multiple responses, where farmers are asked about their veterinary information sources.

Chapter 6 then investigates modelling strategies for a repeated multiple response variable, which also has not been done before. As for multiple response data, GEE and HLP models are two common model strategies, but the increased

number of items of the repeated multiple responses often makes HLP models infeasible. We also propose a new way to estimate the correlation for grouped observations, obtaining more efficient parameter estimates for the GEE method.

Chapter 7 discusses graphical diagnostic methods for GEE extending a univariate cumulative residual process (Lin et al. 2002) to a multivariate process. The methods are applied to the proportional odds model to check the functional form of a covariate, whereas many other (graphical) diagnostic methods check only the overall model adequacy.

The last chapter (Ch. 8) summarises the results of the dissertation and discusses some further research topics for future work.

Chapter 2

The Analysis of Stratified Multiple Responses

2.1 Introduction

In many surveys respondents may select any number of the outcome categories. For instance, in Section 1.1.1 on page 1, we considered the question "What type of contraceptives have you used?" with possible responses (A-oral, B-condom, C-lubricated condom, D-spermicide, and E-diaphragm), where respondents are asked to tick all items that apply. We can cross-classify these counts from a survey that contains a multiple response variable with $J = 5$ items along with a group variable (r levels, e.g. whether a subject had a prior history of urinary tract infection) and a stratification variable (K levels, e.g. several age groups) into an $r \times J \times K$ contingency table. An example due to Bilder and Loughin (2002) is the $2 \times 5 \times 2$ Table 1.1 for 239 sexually active college women. We are interested in the conditional relationship between the type of contraception and prior history of urinary tract infection, given the age group.

Another example comes from a study conducted by Dr. Paul Warren in the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. Six experts (raters) rated 50 utterances by non-native English speakers on a 3-point scale for overall comprehensibility (from "not easy" to "very easy" to understand) and then indicated whether there was a problem for each utterance in each of the 7 items (e.g. pronunciation of consonants, vowel pronunciation, word stress, etc.). These 7 items are the pick any/J variables, where $J = 7$ in this example. Each item can be treated as a binary choice (i.e., it was or was not a problem). The study was interested in evaluating the conditional relationship between the overall rating and the 7 items, given the raters. Table 2.1 comprises 6 separate 3×7 tables ($K = 6$, $r = 3$ and $J = 7$), where the cell counts are dependent across the columns for each table and also dependent across the 6 strata.

Both examples are of stratified multiple response data, yet the observations are not independent across the strata in the second example. Such data occur frequently in health and social sciences and in language studies. To analyse the data we need the complete information on which items have been selected for each of the women (Example 1) or utterances (Example 2). One can express the complete information for each of the respondents using an $r \times 2^J \times K$ contingency table such as Table 2.2, where the columns form the response profile on the J items. In total there are 2^J possible profiles, according to the (yes, no) outcome for the selection of each item. The complete information on each of the 50 utterances on the 6 raters and 7 items can be displayed in a similar fashion.

Cochran-(MH) test statistics determine whether a response variable is independent of another variable given a third variable. Bilder and Loughin (2002) generalised the Cochran test to determine whether the group and pick any/J

Table 2.1: The marginal linguistics data

	Items							Total	Total
	1	2	3	4	5	6	7	responses	utterances
Rater 1									
Rating									
1	8	7	2	2	1	0	1	21	8
2	32	22	7	2	6	0	3	72	32
3	8	1	3	0	0	0	1	13	10
Rater 2									
Rating									
1	10	8	8	4	5	8	0	43	11
2	18	6	10	11	8	11	1	65	19
3	18	9	4	3	8	7	0	49	20
Rater 3									
Rating									
1	7	1	3	0	4	2	0	17	7
2	11	4	6	1	8	4	0	34	13
3	23	7	8	3	13	8	2	64	30
Rater 4									
Rating									
1	2	2	2	2	0	0	0	8	2
2	11	7	2	4	1	1	0	26	12
3	11	6	1	5	0	0	1	24	36
Rater 5									
Rating									
1	1	0	0	0	0	0	0	1	1
2	8	6	5	0	1	1	0	21	23
3	5	11	4	0	1	1	0	22	26
Rater 6									
Rating									
1	14	18	6	14	14	17	0	83	18
2	12	10	1	9	11	9	0	52	14
3	12	14	4	7	9	11	1	58	18

Table 2.2: The complete UTI data

Age ≥ 24																
Contraceptive																
Oral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Condom	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
L. cond.	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Spermicide	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Diaphragm	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
UTI																
No	0	0	0	0	0	0	0	0	0	0	0	0	2	0	4	0
Yes	0	0	0	0	0	0	0	0	2	0	1	1	1	0	0	1
Contraceptive																
Oral	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Condom	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
L. cond.	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Spermicide	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Diaphragm	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
UTI																
No	14	0	1	0	0	0	0	0	1	0	0	0	0	0	2	0
Yes	5	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
Age < 24																
Contraceptive																
Oral	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Condom	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
L. cond.	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Spermicide	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Diaphragm	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
UTI																
No	0	0	1	0	0	0	0	0	2	0	1	0	8	0	18	0
Yes	0	0	1	0	0	0	0	0	14	0	3	0	10	0	12	1
Contraceptive																
Oral	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Condom	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
L. cond.	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Spermicide	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Diaphragm	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
UTI																
No	42	0	1	0	0	0	0	0	1	0	0	0	5	0	6	0
Yes	44	3	0	0	0	0	0	0	15	1	2	0	7	0	3	0

variable (or "items") are marginally independent given a stratification variable This is known as *conditional multiple marginal independence* (CMMI). For the UTI example, they tested whether the contraception practices of women are different based on their urinary tract infection history, controlling for their age group. They used a nonparametric bootstrap method to obtain the p-value of the test. When the group and items are not conditionally marginally independent, it is more interesting to describe how the items depend on the group. Similarly, in the linguistics example we are not interested in the differences between raters, and we focus on describing the conditional relationship between the overall rating and the items, given each rater.

This chapter discusses two approaches to the analysis of such data. The first approach, called the *model-based* approach, treats the J items as a J-dimensional binary response and then uses logit models directly for the marginal distribution of each item. It applies the methodology of generalised estimation equations (GEE) (Liang and Zeger 1986), a multivariate extension of quasi-likelihood methods. The GEE method is the computationally simplest one as we need only to provide the mean-variance relationship and specify the working correlation structure for the J items..

The Mantel-Haenszel type method, called the *non-model-based* approach, is another option. This second approach extends the generalised Mantel-Haenszel (GMH) estimators of Greenland (1989) to make the inference across J items. The MH-type estimators are dually consistent, i.e. consistent under the limiting model I (the "large stratum" limiting model) and model II (the "sparse data" limiting model).

For an ordinary binary response case, it is well known that the MH estimators perform much better than the ML estimators for sparse data (Anderson 1980,

p.244). To make the inference across J items, we derive the dually consistent variance and covariance estimators for the generalised MH estimators. As in the Cochran-Mantel-Haenszel test, generalised MH estimators are used when the conditional associations are not expected to vary drastically among the strata. However, even though the true associations are heterogeneous between strata, the generalised MH estimators often provide a useful descriptive summary if the directions of the associations are the same across strata. When heterogeneity exists, it is always interesting to get the details of the heterogeneity. For the linguistics example, although we want summary information on the conditional relationship between the overall rating and the items given each rater, it is also useful to find out which rater (if any) differs from the others. We use the influence measure of Liu and Wang (2007) to evaluate the heterogeneity among raters.

Section 2.2 introduces the model-based approach using the GEE method, then Section 2.3 shows how the generalised MH estimators apply to multiple responses and gives dually consistent variance and covariance estimators. Section 2.4 provides the data analysis for the two examples. The dually consistent variance and covariance estimators for the generalised MH estimators are applicable only when the strata are independent. When the strata are dependent, as in the linguistics example, it is more realistic to use the bootstrap method to evaluate the variance and covariances of the estimators, because the dually consistent ones are too complicated to derive. Therefore, in Section 2.5, we discuss the simulation results for the performance of the bootstrap method when the data are simulated from various situations. We also compare the relative performances of the GEE and MH methods. Section 2.6 uses an influence measure to analyse the heterogeneity between the strata. The next section (Sec. 2.7) provides a general discussion. Finally, Section 2.8 proves in detail the dual consistency and the formula for

the generalised MH estimator, and justifies the choice of the influence measure. We published Sections 2.2-2.5 and 2.7 in a similar form (Liu and Suesse 2008).

2.2 Model Based Approach

Consider the J items as a J-dimensional binary response. For each item, the response is either "the item is selected" or "the item is not selected". For example, for linguistics data we let $\pi_{x|ak}$ be the probability of having a problem on item x when the utterance is overall rated on level a by rater k. To describe our main course, the conditional relationship between the overall rating and the items given each rater, we use the logit model for the marginal probabilities of each item having the form

$$\log\left(\frac{\pi_{x|ak}}{1-\pi_{x|ak}}\right) = \beta_{ax} + \tau_{xk}\,, \tag{2.1}$$

where $a = 1,\ldots,r$, $x = 1,\ldots,J$, and $k = 1,\ldots,K$. Identifiability requires constraints such as either $\beta_{rx} = 0$ or $\tau_{xK} = 0$ for all x ($K, r \geq 2$). Define $\gamma_{ab}^{x} = \beta_{ax} - \beta_{bx}$. The parameters $\{\gamma_{ab}^{x}\}$ characterise the conditional relationships. For instance, the odds of having a problem on item x when the utterance is overall rated on level a are $\exp(\gamma_{ab}^{x})$ times the odds of having a problem on item x when the utterance is overall rated on level b, given each rater.

We fit the model by applying the GEE methodology; see section 5.2.2 on page 150 for details. Since the GEE method is a multivariate extension of the quasi-likelihood method, we do not need to specify the full joint distribution of the J items. Only the mean-variance relationship and the correlation structure for the J items need to be specified. One can make a "working guess" about the correlation structure of the item responses and then adjust the standard error of

the parameter estimators to reflect what actually occurs in the sample data using a "sandwich" method.

The GEE approach is easy to apply for the UTI data, because the responses are dependent only across the J items and the observations are independent across strata. For the linguistics data it is not clear how the correlation structure can be chosen when the responses are correlated across the J items and also the K strata (raters), although one could always choose an "independent" working correlation structure and use the sandwich standard errors to take into account the empirical situation.

Instead of using the logit model, the conditional associations can also be obtained using a generalised MH-type estimator. Unlike the logit model, the MH-type method cannot be used to select the best model that includes all significant predictors. However, if one is particularly interested in the conditional association between the item and the overall rating given each rater, the MH-type estimators evaluate the association directly. The next section gives the details.

2.3 Non-Model Based Approach

2.3.1 MH Estimators

Let us consider each item separately. For example, consider only item "1" (consonant pronunciation) in Table 2.1. The conditional association between overall rating and "whether there was a consonant pronunciation problem" given the rater can be described using a $3 \times 2 \times 6$ table, where the column variable is "whether there was a consonant pronunciation problem" with two levels (yes, no), the row variable is overall rating (not easy, medium, very easy), and the stratum variable is rater. Suppose we naively treat the 3×2 tables for the 6 raters as independent.

We can use the generalised MH estimators (Greenland 1989) to describe the conditional relationship between the row and column variables. The dual consistency of the estimators (MH and generalised MH) have already been established. However, the standard error and covariance estimates for the estimators based on the naive independence assumption are inappropriate and the (dual) consistency of the (co)variance estimators need to be determined. There are two ways to find proper standard errors and covariance estimates: (1) deriving dually consistent estimators; and (2) using the bootstrap method. We will discuss these in Sections 2.3.2 and 2.3.3.

For a general $r \times J \times K$ table, let $X_{x|ak}$ denote the number of utterances having a problem on item x rated by the kth rater (stratum) with the overall rating (row) a. The notation n_{ak} denotes the total number of utterances in the ath row and the kth stratum. Let $N_k = n_{1k} + \cdots + n_{rk}$. For convenience, we also let $\bar{\pi}_{x|ak} = 1 - \pi_{x|ak}$ and $\bar{X}_{x|ak} = n_{ik} - X_{x|ak}$. Define a common odds ratio for rows a and b as

$$\Psi^x_{ab} = \Psi^x_{abk} = \frac{\pi_{x|ak}\bar{\pi}_{x|bk}}{\bar{\pi}_{x|ak}\pi_{x|bk}} \quad x = 1, \ldots, J, \;\; a, b = 1, \ldots, r \; (a \neq b), \tag{2.2}$$

for all k. Ψ^x_{ab} is the ratio of the odds of having a problem on item x for utterances overall rated a to the odds of having a problem on item x for utterances overall rated b, given any stratum. The ordinary MH estimator is

$$\hat{\Psi}^x_{ab} = \frac{C_{x|ab}}{C_{x|ba}}\,, \tag{2.3}$$

where $C_{x|ab} = \sum_{k=1}^{K} c_{x|abk}$ with $c_{x|abk} = X_{x|ak}\bar{X}_{x|bk}/N_k$. Greenland (1989) introduced the generalised MH estimator of $\log \Psi^x_{ab}$

$$\bar{L}^x_{ab} = (L^x_{a+} - L^x_{b+})/r\,, \tag{2.4}$$

where $L^x_{ab} = \log \hat{\Psi}^x_{ab}$ and the subscript "+" indicates summation over that subscript. When the row variable has only two levels ($r = 2$) as for the UTI example, we can use Ψ^x_{12} to describe the conditional row effect on selecting item x. For $r = 2$ the generalised MH estimator of $\log \Psi^x_{12}$ simplifies to the ordinary MH estimator

$$L^x_{12} = \log \left(\frac{\sum_{k=1}^{K} X_{x|1k} \bar{X}_{x|2k} / N_k}{\sum_{k=1}^{K} X_{x|2k} \bar{X}_{x|1k} / N_k} \right). \tag{2.5}$$

2.3.2 Dually Consistent Variance and Covariance Estimators

When the strata are independent (as in the UTI example), we can derive the dually consistent variance and covariance estimators for the generalised MH estimators. If one is only interested in a particular item, say x ($x \in \{1, \ldots, J\}$), the dually consistent variance and covariance estimators for $\{\bar{L}^x_{ab}, \forall\, a \neq b\}$ come directly from the work by Greenland (1989). However, one might be interested in comparing the conditional association across items. For instance, in the UTI example one might be interested in comparing the UTI effects of the contraceptive methods "oral" and "condom". The covariance estimator between $\bar{L}^x_{ab}$ and $\bar{L}^y_{ab}$ is desirable for $x \neq y$ ($x, y \in \{1, \ldots, J\}$). The way to derive the dually consistent estimator for it is more complicated than the case considering only a fixed item, because $X_{x|ak}$ and $X_{y|ak}$ are correlated for all a and k. That is, the numbers of women who used contraceptive methods x and y are not independent. To find the dually consistent covariance estimator, we need to consider up to the fourth moment of the X's and the pairwise counts for the two items.

Define pairwise probabilities for items x and y ($x, y \in \{1, \ldots, J\}$) as $\pi^{st}_{xy|ak}$ with $s, t \in \{0, 1\}$, where $(0, 1)$ is the (no, yes) outcome for the selection of each item. Then $\pi^{st}_{xy|ak}$ is the probability of observing the pairwise outcome (s, t) for items x and y. For instance, the notation $\pi^{11}_{xy|ak}$ represents the probability that a subject,

who is in row a and stratum k, selects both items x and y. Define similarly the pairwise observations as $\{X^{st}_{xy|ak}\}$. We assume $\mathbf{X}_{xy|ik} = (X^{00}_{xy|ak}, X^{01}_{xy|ak}, X^{10}_{xy|ak}, X^{11}_{xy|ak})$ follows a multinomial distribution with parameters n_{ak} and $\boldsymbol{\pi}_{xy|ak} = (\pi^{00}_{xy|ak}, \pi^{01}_{xy|ak}, \pi^{10}_{xy|ak}, \pi^{11}_{xy|ak})$ with $\pi^{00}_{xy|ak} + \pi^{01}_{xy|ak} + \pi^{10}_{xy|ak} + \pi^{11}_{xy|ak} = 1$. The marginal probabilities can be computed from the pairwise probabilities by $\pi_{x|ak} = \pi^{10}_{xy|ak} + \pi^{11}_{xy|ak}$ and $\pi_{y|ak} = \pi^{01}_{xy|ak} + \pi^{11}_{xy|ak}$.

First we consider the fixed item x. Define $h_{x|ab} = (X_{x|ak} + \bar{X}_{x|bk})/N_k$. Let $\mathbb{E}$ denote the standard expectation, Var and Cov the standard variance and covariance. Greenland (1989) derived the following estimators

$$U_{x|abb} := \widehat{\text{Var}}(L^x_{ab}) \qquad\qquad U_{x|abc} := \widehat{\text{Cov}}(L^x_{ab}, L^x_{ac}) \tag{2.6}$$

with

$$\begin{aligned} U_{x|abb} &= \frac{\sum_k c_{x|ab} h_{x|ab}}{2C^2_{|xab}} + \frac{\sum_k c_{x|ba} h_{x|ab} + c_{x|ab} h_{x|ba}}{2C_{x|ab}C_{x|ba}} + \frac{\sum_k c_{x|ba} h_{x|ba}}{2C^2_{x|ba}} \\ U_{x|abc} &= \frac{\sum_k X_{x|a}\bar{X}_{x|b}\bar{X}_{x|c}/N_k^2}{3C_{x|ab}C_{x|ac}} + \frac{\sum_k n_a \bar{X}_{x|b} X_{x|c}/N_k^2}{3C_{x|ab}C_{x|ca}} \\ &\quad + \frac{\sum_k n_a X_{x|b}\bar{X}_{x|c}/N_k^2}{3C_{x|ba}C_{x|ac}} + \frac{\sum_k \bar{X}_{x|a} X_{x|b} X_{x|c}/N_k^2}{3C_{x|ba}C_{x|ca}}. \end{aligned}$$

Please note, subscript c refers to a row (as the indices a and b do). For convenience, we often suppress subscripts x and k. For instance, $c_{ab} = c_{x|abk}$, $X_a = X_{x|ak}$, and $n_a = n_{ak}$.

Because $\bar{L}^x_{ab}$ is a linear combination of $\{L^x_{ab}\}$, $\widehat{\text{Cov}}(\bar{L}^x_{ab}, \bar{L}^x_{cd})$ can be expressed as follows in terms of $U_{x|ab}$ and $U_{x|abc}$

$$\widehat{\text{Cov}}(\bar{L}^x_{ab}, \bar{L}^x_{cd}) = (U^+_{x|ac} - U^+_{x|ad} - U^+_{x|bc} + U^+_{x|bd})/r^2 \tag{2.7}$$

with

$$U^{+}_{x|ab} = \begin{cases} U_{x|a++} = \sum_{i,h} U_{x|aih} & \text{for } a = b \\ U_{x|+ab} - U_{x|ab+} - U_{x|ba+} + U_{x|ab} = U^{+}_{x|ba} & \text{for } a \neq b \end{cases}$$

The subscript "+" denotes summation over that subscript. Note that setting $c = a$, $d = b$ yields $\widehat{\text{Var}}(\bar{L}^x_{ab})$ and setting $c = a$, $d = c$ yields $\widehat{\text{Cov}}(\bar{L}^x_{ab}, \bar{L}^x_{ac})$.

Next, we make the inference across two different items. For instance, consider the covariance between $\bar{L}^x_{ab}$ and $\bar{L}^y_{cd}$. We propose the following dually consistent covariance estimators

$$U_{xy|abb} := \widehat{\text{Cov}}(L^x_{ab}, L^y_{ab}) = \frac{\hat{D}^{11}_{ab}}{C_{x|ab}C_{y|ab}} - \frac{\hat{D}^{01}_{ab}}{C_{x|ba}C_{y|ab}} - \frac{\hat{D}^{10}_{ab}}{C_{x|ab}C_{y|ba}} + \frac{\hat{D}^{00}_{ab}}{C_{x|ba}C_{y|ba}} \quad (2.8)$$

$$U_{xy|abc} := \widehat{\text{Cov}}(L^x_{ab}, L^y_{ac}) = \frac{\hat{D}^{11}_{abc}}{C_{x|ab}C_{y|ac}} - \frac{\hat{D}^{01}_{abc}}{C_{x|ba}C_{y|ac}} - \frac{\hat{D}^{10}_{abc}}{C_{x|ab}C_{y|ca}} + \frac{\hat{D}^{00}_{abc}}{C_{x|ba}C_{y|ca}} \quad (2.9)$$

with $\hat{D} = \sum_k \hat{d}_k$,

$$\hat{d}^{st}_{ab} = \frac{1}{N^2_k}\{X^s_{x|a}X^t_{y|a}X^{\bar{s}\bar{t}}_{xy|b} + X^{st}_{xy|a}X^{\bar{s}}_{x|b}X^{\bar{t}}_{y|b} - X^{st}_{xy|a}X^{\bar{s}\bar{t}}_{xy|b}\} \quad (2.10)$$

and

$$\hat{d}^{st}_{abc} = \frac{1}{N^2_k}X^{st}_{xy|a}X^{\bar{s}}_{x|b}X^{\bar{t}}_{y|c}, \quad (2.11)$$

where we set $\bar{s} := 1 - s$ and use the convenient notation $X^1_{x|a} := X^s_{x|a}$ and $X^0_{x|a} := \bar{X}^s_{x|a}$, for example for the pairwise counts $X^{\bar{1}\bar{1}}_{xy|a} = X^{00}_{xy|a}$. We see that estimator $U_{xy|abb}$ has a similar form as $U_{x|abb}$ and $U_{xy|abc}$ is similar to $U_{x|abc}$, specially when comparing the denominators. For $a \neq b$ and $c \neq d$, $U_{xy|abcd}$ does not need to be defined, because $\text{Cov}(L^x_{ab}, L^y_{cd}) = 0$ owing to the independence of rows but $\text{Cov}(L^x_{ab}, L^y_{ac}) \neq 0$. The estimator $U_{xy|abb}$ is invariant under interchange of items (x and y) or rows (a and b). Note that $U_{x|abb} = U_{x|baa}$, because $L^x_{ab} = -L^x_{ba}$. Also, $U_{x|abc} = U_{x|acb}$ by definition. However, $U_{xy|abb} \neq U_{xy|baa}$, but $U_{xy|abb} = U_{yx|baa}$ by

definition.

Again, since $\bar{L}^x_{ab}$ (or $\bar{L}^y_{ab}$) is a linear combination of $\{L^x_{ab}\}$ (or $\{L^y_{ab}\}$), we can derive covariance estimators for $(\bar{L}^x_{ab}, \bar{L}^y_{cd})$, which can be expressed as follows:

$$\widehat{\text{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{cd}) = \frac{1}{r^2}\{U^+_{xy|ac} - U^+_{xy|ad} - U^+_{xy|bc} + U^+_{xy|bd}\} \tag{2.12}$$

with

$$U^+_{xy|ac} = \begin{cases} U_{xy|a++} = \sum_{i,h} \text{Cov}(L^x_{ai}, L^y_{ah}) & \text{for } a = c \\ U_{xy|+ac} - U_{xy|a+c} - U_{xy|ca+} + U_{xy|ac} & \text{for } a \neq c \end{cases}$$

For non-distinct indices a, b, c, d we have

$$\widehat{\text{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{ac}) = \frac{1}{r^2}\{U^+_{xy|a++} - U^+_{xy|ac} - U^+_{xy|ba} + U^+_{xy|bc}\}$$

and

$$\widehat{\text{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{ab}) = \frac{1}{r^2}\{U_{xy|a++} - U^+_{xy|ab} - U^+_{xy|ba} + U_{xy|b++}\}.$$

Note that the contruction of $\widehat{\text{Cov}}(L^x_{ab}, L^y_{cd})$ is very similar to that of $\widehat{\text{Cov}}(L^x_{ab}, L^x_{cd})$, since estimator $\bar{L}$ is a linear combination of the L's for both situations. In the last section (Subsection 2.8.1 on page 72) we prove the dual consistency of these new covariance estimators. We shall refer later to "formulae" variance and covariance estimators, meaning Greenland's dually consistent variance $\widehat{\text{Var}}(\bar{L}^x_{ab})$ in (2.7) and the dually consistent covariance $\widehat{\text{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{cd})$ in (2.12).

When the strata are not independent (as in the linguistic example), it is even more complicated to derive the dually consistent variance and covariance estimators, because the X's are correlated across not only items but also strata. For instance, for $K = 6$ we need to consider up to the 24^{th} ($4 \times 6 = 24$) moment of the X's. Because of this complexity, Section 3.2 provides a realistic way to find

estimates by applying the nonparametric bootstrap method.

2.3.3 Bootstrap Estimates of Variance and Covariance

The nonparametric bootstrap method (Efron and Tibshirani 1993) was conducted by randomly selecting subjects with replacement from the original data. For instance, for the UTI data, we resample N_k women with replacement from the k^{th} stratum, where $k = 1, 2$. Similarly, for the Linguistics example, we resample 50 utterances with replacement and cross-classify the data into a $3 \times 7 \times 6$ table. For each resampled data set, the size of each stratum is the same as before. We take B resamples and then for each resample we calculate the generalised MH estimates $\{\bar{L}^x_{ab},\, x = 1, \ldots, J,\, a \neq b = 1, \ldots, r\}$. The bootstrap estimate of the standard error of $\bar{L}^x_{ab}$ is the standard deviation of the bootstrap replicates,

$$\text{s.e. for } \bar{L}^x_{ab} = \sqrt{\frac{\sum_{s=1}^{B}\left(\bar{L}^x_{ab,s} - \sum_{s=1}^{B}\bar{L}^x_{ab,s}/B\right)^2}{B-1}},$$

where $\bar{L}^x_{ab,s}$ is the generalised MH estimate $\bar{L}^x_{ab}$ for the s^{th} bootstrap resample. Similarly, the bootstrap estimate of the covariance of $\bar{L}^x_{ab}$ and $\bar{L}^y_{cd}$ is

$$\hat{\text{cov}}(\bar{L}^x_{ab},\, \bar{L}^y_{cd}) = \frac{\sum_{s=1}^{B}\left(\bar{L}^x_{ab,s} - \sum_{s=1}^{B}\bar{L}^x_{ab,s}/B\right)\left(\bar{L}^y_{cd,s} - \sum_{s=1}^{B}\bar{L}^y_{cd,s}/B\right)}{B-1}.$$

Later, we will refer to this simply as the "bootstrap" estimate.

Based on the unconditional coefficient of variation, Efron and Tibshirani (1993) suggest that as little as 25 bootstrap samples or replicates are sufficient in obtaining reasonable results for variance estimation. However Booth and Sakar (1998) investigated the precision of the bootstrap variance based on a conditional analysis and conclude that a much higher number of replicates is needed. In the

following, we will use $B = 50,000$ for the data analysis of the examples to obtain sufficient precision.

There are other more efficient bootstrap methods (Shao and Tu 1995), such as bootstrapping based on the studentised residuals or the "bootstrap accelerated bias-corrected percentile"(BC_a). The first method resamples the residuals and the second is based on a transformation of a random variable, for which the distribution is known. However how do we define residuals for the MH estimator and how do we check the assumptions for BC_a? We do not want to investigate the performance and applicability of these and other methods here. The practitioner must keep in mind that there might be other bootstrap methods, which perform better. A careful investigation might be subject to future research.

Remark 2.3.1. *Estimation and Confidence Intervals for the Odds Ratio.* We could estimate Ψ by $\widehat{\Psi}_{xy} \equiv \exp(L)$ but because of efficiency advantages of $\bar{L}$ over L (Greenland 1989), we prefer to estimate Ψ by $\exp(\bar{L})$. The covariances for all indices x, y referring to items and a, b, c, d referring to rows can be computed by

$$\widehat{\text{Cov}}(\exp(\bar{L}^x_{ab}), \exp(\bar{L}^y_{cd})) = \exp(\bar{L}^x_{ab} + \bar{L}^y_{cd})\widehat{\text{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{cd})$$

because

$$\text{Cov}^a(\exp(\bar{L}^x_{ab}), \exp(\bar{L}^y_{cd})) = \exp(\log \Psi^x_{ab}) \exp(\log \Psi^y_{cd})\text{Cov}^a(\bar{L}^x_{ab}, \bar{L}^y_{cd})$$

by the Delta method, where Cov^a stands for the asymptotic covariance. However, a confidence interval is best constructed in the log-scale (Emerson 1994), because of the log-scale's symmetry. An approximate 95% confidence interval for the log-odds is

$$\bar{L} - 1.96\sqrt{\widehat{\text{Var}}(\bar{L})} \leq \log \Psi \leq \bar{L} + 1.96\sqrt{\widehat{\text{Var}}(\bar{L})}$$

and an approximate 95% confidence interval for the odds ratio is

$$\exp[\bar{L} - 1.96\sqrt{\widehat{\text{Var}}(\bar{L})}] \leq \Psi \leq \exp[\bar{L} + 1.96\sqrt{\widehat{\text{Var}}(\bar{L})}].$$

2.4 Examples

For the UTI example, the model-based (GEE) approach gives $\{\hat{\gamma}_{12}^x,\ x = 1, \ldots, 5\} = \{0.12\ ,\ -0.52,\ 0.71,\ 0.65,\ -8.96\}$ with sandwich standard errors $\{0.28,\ 0.27,\ 0.28,\ 0.31,\ 1.10\}$ using an exchangeable correlation structure. Alternatively, the non-model-based (MH) approach gives $\{L_{12}^x,\ x = 1, \ldots, 5\}$=$\{0.12\ ,\ -0.52,\ 0.71,\ 0.64,\ -2.57\}$ with standard errors $\{0.28, 0.26, 0.28, 0.31, 1.41\}$ by applying formula (2.7). Choosing $B = 50,000$, the corresponding bootstrap standard errors are $\{0.28,\ 0.26,\ 0.28,\ 0.32,\ 0.39\}$. For instance, for the first item (oral contraceptive), the odds of having used the oral contraceptive for women without a prior history of UTI are estimated to be $\exp(0.12) = 1.13$ times the odds for women with a prior history of UTI, given each age group.

The two approaches have similar results, except for the last item ("Diaphragm"), because our data have no women without a prior history of urinary tract infection who use diaphragms. In Table 1.1, the cell count for row 1 and column 5 is zero for both age groups. The GEE estimation routine fails to provide sandwich standard errors. Similarly, the MH estimate L_{12}^5 is undefined. To overcome this problem in the model-based approach, we add to the data set a pseudo-subject with no UTI history who used a diaphragm.

The model (2.1) is fitted by giving the pseudo-subject a small weight (say, 10^{-3}). For the non-model-based approach, one way to get an amended estimator is by adding 0.5 to each cell as suggested by Agresti (2002, p.71) for the ordinary

Table 2.3: The bootstrap with $B = 50,000$ and formulae (in parentheses) variance and covariance estimates of $\{L_{12}^x,\ x = 1,\ldots,5\}$, $10\times$ co-/variance for the data in Table 1.1 (UTI data)

$\mathrm{Cov}(L_{12}^x, L_{12}^y)$					
			y		
x	1	2	3	4	5
1	$0.79(0.76)$	$-0.50(-0.48)$	$-0.45(-0.42)$	$-0.48(-0.02)$	$0.11(NA)$
2	$-0.50(-0.48)$	$0.68(0.70)$	$0.51(-0.37)$	$0.45(0.60)$	$-0.07(NA)$
3	$-0.45(-0.42)$	$0.51(-0.37)$	$0.81(0.80)$	$0.51(0.44)$	$-0.06(NA)$
4	$-0.48(-0.02)$	$0.45(0.60)$	$0.51(0.44)$	$1.04(0.94)$	$-0.012(NA)$
5	$0.11(NA)$	$-0.07(NA)$	$-0.06(NA)$	$-0.12(NA)$	$1.52(19.94)$

items: 1-oral, 2-condom, 3-l.condom, 4-spermicide, 5-diaphragm

NA: not applicable

Table 2.4: 95% confidence intervals for $\log\Psi_{12}^x - \log\Psi_{12}^y$ for the data in Table 1.1 (UTI data) based on formulae (lower left half) and bootstrap with $B = 50,000$ (upper right half) (co)variance estimates

			y		
	1	2	3	4	5
x	oral	condom	l.condom	spermicide	diaphragm
1 oral		$(-1.6140,\ 0.3342)$	$(-0.3876,\ 1.5724)$	$(-0.5117,\ 1.5589)$	$(-3.5851,\ -1.7931)$
2 condom	$(-1.6055,\ 0.3257)$		$(0.8074,\ 1.6572)$	$(0.6022,\ 1.7248)$	$(-2.9973,\ -1.1011)$
3 l.condom	$(-0.3684,\ 1.5532)$	$(0.3034,\ 2.1613)$		$(-0.6335,\ 0.4959)$	$(-4.2517,\ -2.3113)$
4 spermicide	$(-0.2976,\ 1.3375)$	$(0.7488,\ 1.5711)$	$(-0.6448,\ 0.5047)$		$(-4.2498,\ -2.1756)$
5 diaphragm	(NA)	(NA)	(NA)	(NA)	

NA: not applicable

Table 2.5: The generalised MH estimates and their bootstrap standard errors with $B = 50,000$ (in parentheses) for the data in Table 2.1 (Linguistic data)

	item j						
	1 pronunciation of consonants	2 pronunciation of vowels	3 word stress	4 sentence stress	5 rhythm	6 intonation	7 rate
$\bar{L}^j_{12}$	-0.00	1.19	0.70	0.28	-0.10	0.88	-0.39
	(0.81)	(0.50)	(0.53)	(0.40)	(0.47)	(0.50)	(1.07)
$\bar{L}^j_{13}$	1.34	1.47	1.21	1.49	0.73	1.36	-1.23
	(0.73)	(0.48)	(0.58)	(0.49)	(0.44)	(0.45)	(1.17)
$\bar{L}^j_{23}$	1.34	0.27	0.52	1.20	0.83	0.48	-0.84
	(0.52)	(0.30)	(0.47)	(0.50)	(0.50)	(0.43)	(1.35)

odds ratio estimator. The cell counts for a stratum having only a few observations are usually small. If we add 0.5 to a small cell count, it could easily influence and weaken the association. In order not to smooth the data too much, we add 0.5 to each cell for the stratum with largest size. For instance, because the stratum of Age<24 contains the greater number of observations, we add 0.5 to each cell in that stratum. The estimate of the odds ratio for the last item ("Diaphragm") is not stable under either approach. In summary, the conditional UTI effects are significant for the contraceptives "condom", "lubricated condom", and "spermicide" at a 5% significance level.

Table 2.3 gives the bootstrap with $B = 50,000$ and formulae (shown in parentheses) variance and covariance estimates using equations (2.6) and (2.8) for $\{L^x_{12}, x = 1,\ldots,5\}$. Table 2.4 shows all multiple comparisons of the conditional UTI effects for any two items. For instance, comparing the UTI effects for the contraceptives "oral" and "lubricated condom", a 95% confidence interval for $\log \Psi^1_{12} - \log \Psi^3_{12}$ is $(-0.39, 1.57)$. Owing to the sampling zero for item 5 (Diaphragm), a few consistent covariance estimators involving L^5_{12} are not applicable. Consequently,

the confidence intervals based on the formulae are not applicable for item 5. Alternatively, one can choose to amend the pairwise observations to obtain rough estimates for them.

In the linguistics example, the GEE approach fails to give the sandwich standard errors for the model (2.1). Instead we fit a parsimonious model that replaces τ_{xk} by $\tau_x + \alpha_k$. However, the generalised MH estimator works for the general model (2.1). By comparing overall rating levels 1 and 2, the MH estimates $\{\bar{L}^x_{12},\, x = 1, \ldots, 7\}$ are $\{-0.00,\, 1.19,\, 0.70,\, 0.28,\, -0.10,\, 0.88,\, -0.39\}$ with the bootstrap standard errors with $B = 50,000$ of $\{0.81,\, 0.50,\, 0.53,\, 0.40,\, 0.47,\, 0.50,\, 1.07\}$. Comparing rating levels 1 and 3, the MH estimates $\{\bar{L}^x_{13},\, x = 1, \ldots, 7\}$ is $\{1.34,\, 1.47,\, 1.21,\, 1.49,\, 0.73,\, 1.36,\, -1.23\}$ with the bootstrap standard errors with $B = 50,000$ of $\{0.73,\, 0.48,\, 0.58,\, 0.49,\, 0.44,\, 0.45,\, 1.17\}$. There are no significant differences between rating levels 1 and 2 for any item except for item 2 (pronunciation of vowels), given each rater. However, the differences between rating levels 1 and 3 are significant, given each rater, for all items except items 1, 5, and 7. Table 2.5 shows the generalised MH estimates and their bootstrap standard errors. Similarly, the bootstrap variance and covariance estimates can be calculated. In this example the formulae (co)variance estimators are not appropriate, because this data set has dependent strata.

Although the GEE method (the model-based approach) uses a more parsimonious model than the MH method (the non-model-based approach), both methods give similar results in terms of the significance. For instance, the GEE estimates for $\{\log \Psi^x_{13}, x = 1, \ldots, 7\}$ are $\{1.34,\, 1.32,\, 0.83,\, 1.29,\, 0.76,\, 1.24,\, -1.11\}$ with the sandwich standard errors $\{0.67,\, 0.39,\, 0.53,\, 0.38,\, 0.33,\, 0.32,\, 1.19\}$.

2.5 Simulation Study

In the simulation study we evaluate the performance of the model-based (GEE) and non-model-based (MH) estimators for the odds ratio and their (co)variances estimators. The simulation study consists of two main cases. One case assumes that the strata are independent as in the UTI example. The other case allows dependency between strata as in the linguistics example. In case 1 the scenarios range from ones for which the limiting model I should work well to ones for which the asymptotic model II seems more appropriate. In case 2 the situation varies according to the degree of the dependency between strata.

For the model-based estimators (GEE), we use R (*Team R Development Core, A Language and Environment for Statistical Computing* 2006) and its package geepack (Yan 2004, Yan and Fine 2004) for fitting. We always assume an exchangeable correlation structure to obtain the estimates $\{\hat{\gamma}^x_{ab}; a \neq b; a, b = 1, \ldots, r; x = 1, \ldots, J\}$. We automatically obtain the robust (or sandwich) and naive (co)variances as a by-product of the fitting algorithm. The robust covariance matrix is also consistent, when the working correlation structure does not match the true correlation structure, which is in contrast to the naive covariance. For further details of GEE see Section 5.2.2 on page 150.

For the non-model-based method (MH) we compute $\{\bar{L}^x_{ab};\ a \neq b; a, b = 1, \ldots, r; x = 1, \ldots, J\}$ and its bootstrap and formulae (co)variances.

Independent Strata For simplicity we let $r = 2$ and use a constant odds ratio for every item, i.e., $\Psi^x_{12} = \Psi$ for all $x = 1, \ldots, J$. We also set the marginal probabilities $\pi_{x|1k}$ to be 0.5 for all items $x = 1, \ldots, c$ and strata $k = 1, \ldots, K$. The marginal probabilities $\{\pi_{x|2k}\}$ are computed from the given common odds ratio Ψ. Let Y_x indicate whether a subject selects item x: given a and k, if a subject selects item x

then $Y_x = 1$ otherwise $Y_x = 0$. The pairwise dependency between items x and y is denoted using an odds ratio θ_{xy} as

$$\theta_{xy|ik} = \frac{P(Y_x = 1, Y_y = 1|ak)P(Y_x = 0, Y_y = 0|ak)}{P(Y_x = 0, Y_y = 1|ak)P(Y_x = 1, Y_y = 0|ak)}, \tag{2.13}$$

where $x \neq y = 1, \ldots, J$.

Let $\pi_{xy|ak} := \pi^{11}_{xy|ak} = P(Y_x = 1, Y_y = 1|ak)$, then $P(Y_x = 1, Y_y = 0|ak) = \pi_{x|ak} - \pi_{xy|ak}$, $P(Y_x = 0, Y_y = 1|ak) = \pi_{y|ak} - \pi_{xy|ak}$ and $P(Y_x = 0, Y_y = 0|ak) = 1 - \pi_{x|ak} - \pi_{y|ak} + \pi_{xy|ak}$ with $\pi_{xy|ak}$ satisfying

$$\min(0, \pi_x + \pi_y - 1) \leq \pi_{xy} \leq \max(\pi_x, \pi_y). \tag{2.14}$$

For given θ_{xy}, π_x and π_y, we can compute the unique solution π_{xy} of the quadratic

$$(\theta_{xy} - 1)\pi^2_{xy} - \pi_{xy}[1 + (\theta_{xy} - 1)(\pi_x + \pi_y)] + \theta_{xy}\pi_x\pi_y = 0 \tag{2.15}$$

satisfying (2.14). It follows that the complete pairwise distribution given by pairwise probabilities π^{11}_{xy}, π^{01}_{xy}, π^{10}_{xy} and π^{00}_{xy} is completely specified by θ_{xy} (respectively π_{xy}), π_x and π_y.

Then the 2^J joint probabilities $\mathbf{P}_{\mathbf{Y}|ak} = \{P(Y_1 = s_1, \ldots, Y_J = s_J|ak), s_x = 0, 1; x = 1, \ldots, J\}$ in the complete table (as in Table 2.2) can be computed from the probabilities $\{\pi_{x|ak}, x = 1, \ldots, J\}$ and $\{\pi_{xy|ak}, x \neq y = 1, \ldots, J\}$, if a feasible solution exists (Lee 1993). Usually there are many solutions for the 2^J ($J > 2$) joint probabilities, but for some configurations of $\{\pi_{x|ak}, x = 1, \ldots, J\}$ and $\{\pi_{xy|ak}, x \neq y = 1, \ldots, J\}$, there is no feasible solution. For example: For $J = 3$ and given $\{\pi_{1|ak}, \pi_{2|ak}, \pi_{3|ak}\}$, the parameters θ_{12} and θ_{13} can be chosen arbitrarily determining the pairwise probabilities for the pairs of items $(1, 2)$ and $(1, 3)$. However, they

also constrain the pairwise probabilities for the pair of items $(2, 3)$. Therefore θ_{23} cannot be chosen arbitrarily.

There are several approaches to computing such a solution of the joint probabilities for given pairwise and marginal probabilities. One approach is to use linear programming. Another is applying the iterative proportional fitting (IPF) algorithm as described by Gange (1995). Let $\mathbf{P}^j_{\mathbf{Y}|ak}(= \mathbf{P}_{\mathbf{Y}|ak})$ denote the joint probabilities of the generic jth step for group i and stratum k and $\mathbf{P}^j_{\mathbf{Y}} = (\mathbf{P}^j_{\mathbf{Y}|11}, \ldots, \mathbf{P}^j_{\mathbf{Y}|rK})$. Set $\mathbf{P}^0 = \mathbf{1}$. We can collapse $\mathbf{P}^j$ into marginal probabilities $\{P(Y_x = 1|ak)^j; a = 1, \ldots, r; k = 1, \ldots, K; x = 1, \ldots, J\}$ and pairwise probabilities $\{P(Y_x = 1, Y_y = 1|ak)^j; a = 1, \ldots, r; k = 1, \ldots, K; x, y = 1, \ldots, J\}$. The aim is to find a solution $\mathbf{P}^j$ satisfying $P(Y_x = 1|ak)^j = \pi_{x|ak}$, $P(Y_x = 1, Y_y = 1|ak)^j = \pi_{xy|ak}$ and $P^j_{0|ak} := \sum_{s_1,\ldots,s_J=0,1} P(Y_1 = s_1, \ldots, Y_J = s_J|ak)^j = 1$. The generic jth step of the IPF algorithm uses the following formulae

$$\begin{aligned}
\mathbf{P}^j_{\mathbf{Y}} &= \mathbf{P}^{j-1}_{\mathbf{Y}} \frac{\pi_{x|ak}}{P(Y_x = 1|ak)^{j-1}} \\
\mathbf{P}^j_{\mathbf{Y}} &= \mathbf{P}^{j-1}_{\mathbf{Y}} \frac{\pi_{xy|ak}}{P(Y_x = 1, Y_y = 1|ak)^{j-1}} \\
\mathbf{P}^j_{\mathbf{Y}} &= \mathbf{P}^{j-1}_{\mathbf{Y}} \frac{1}{P_0^{j-1}} \qquad (2.16) \\
&\forall a = 1, \ldots, r; k = 1, \ldots, K; x, y = 1, \ldots, J.
\end{aligned}$$

The generation of the joint probabilities subject to either $\{\pi_{x|ak}, x = 1, \ldots, J\}$ and $\{\pi_{xy|ak},\ x, y = 1, \ldots, J\}$ or $\{\pi_{x|ak}, x = 1, \ldots, J\}$ and $\{\theta_{xy|ak}, x, y = 1, \ldots, J\}$ is analogous to the one applied by Bilder et al. (2000). We prefer IPF over linear programming because it generates strictly positive (> 0) joint probabilities (assuming such a solution exists), in contrast to linear programming, which might produce zero joint probabilities. It seems more plausible, since none of the 2^J

binary sequences is theoretically excluded from the data generation process.

Again for simplicity we let $J = 2$. The dependency between items is assigned by the odds ratio $\theta = \theta_{12}$. We draw N_k samples independently from either row 1 or row 2 with equal probabilities for stratum k and set $N_1 = \cdots = N_K$. Given the randomly chosen row a and stratum k, a sample (consisting of binary sequences of length J) is drawn from the joint distribution $P_{\mathbf{Y}|ak}$. In case 1, we simulate $n = 20000$ datasets based on the joint distributions $\{P_{\mathbf{Y}|ak}\}$ under a variety of configurations. For the bootstrap method, we use the number of bootstrap resamples as $B = 400$. Note that we did not compute the model or non-model-based estimators, when data amendment was required due to the sampling zero problem.

Let the estimator of a parameter of interest δ, e.g. $\log \Psi^2_{12}$, be denoted by $\hat{\delta}$ and let $\hat{\delta}_j$ be the value of the estimator for the jth simulated data set ($j = 1, \ldots, n$). Then the empirical (or sample) mean is defined as $\bar{\delta} = \frac{1}{n}\sum_{j=1}^{n} \hat{\delta}_j$, the empirical variance as $\frac{1}{n}\sum_{j=1}^{n}(\hat{\delta}_j - \bar{\delta})^2$, the mean squared error (mse) as $\frac{1}{n}\sum_{j=1}^{n}(\hat{\delta}_j - \delta)^2$ and the empirical covariance between estimators $\hat{\delta}$ and $\hat{\epsilon}$ as $\frac{1}{n}\sum_{j=1}^{n}(\hat{\delta}_j - \bar{\delta})(\hat{\epsilon}_j - \bar{\epsilon})$.

The second column of Table 2.6 shows the sample means for the generalised MH estimates (L^1_{12}, L^2_{12}) in the first row, and the sample means for the GEE estimates ($\hat{\gamma}^1_{12}, \hat{\gamma}^2_{12}$) in the second row over $n = 20000$ simulations for various scenarios given in the first column. The third column shows the corresponding mean squared errors. We investigate: (1) The performance of the MH (L's) and the GEE ($\hat{\gamma}$'s) estimators by comparing their sample means and the mean square errors (mse), (2) the performance of the MH (co)variance estimators for the formulae and bootstrap methods, and (3) the performance of GEE (co)variance estimators for the robust and naive methods. To compare the performance of the (co)variance estimators, we calculate the "empirical" (co)variances over 20000

simulations.

For the non-model-based approach, we denote the sample mean for formulae (co)variances by formulae$_{MH}$, and the bootstrap (co)variances by BT$_{MH}$. The empirical (co)variance is denoted by emp$_{MH}$. Similarly we denote for the model-based approach the empirical (co)variances by emp$_{GEE}$, the mean of the robust and naive (co)variances by robust$_{GEE}$ and naive$_{GEE}$, respectively. Each entry of columns 4-6 in Table 2.6 consists of three terms. The first two are the variances of the log odds ratio (L's or $\hat{\gamma}$'s) for items 1 and 2, and the third is the covariance of the log odds ratios between items 1 and 2. The first column shows the configuration of parameters K, N_k, Ψ, θ, and the number in parentheses shows the number of samples which were not included in the simulation study due to the sampling zero problem. The total number of simulated samples involved is: $20000 -$ (this number).

Dependent Strata In case 2 we let $r = J = 2$. Unlike case 1, there is some degree of dependency between strata (or raters in the Linguistics example). We introduce another two parameters, Λ_{kl} and $\Gamma_{xy,kl}$, to describe the dependencies between items and between raters. Let Z_k be whether rater k assigns an overall rating 1. If it is a "yes", then $Z_k = 1$; otherwise $Z_k = 0$. Similarly, let $W_{j,k}$ be whether rater k selects item j. If rater k selects item j, then $W_{j,k} = 1$; otherwise $W_{j,k} = 0$. The parameters Λ_{kl} and $\Gamma_{xy,kl}$ are defined as

$$\begin{aligned}
\Lambda_{kl} &= \frac{P(Z_k = 1, Z_l = 1)P(Z_k = 0, Z_l = 0)}{P(Z_k = 0, Z_l = 1)P(Z_k = 1, Z_l = 0)}, \quad k \neq l = 1, \ldots, K; \\
\Gamma_{xy,kl} &= \frac{P(W_{x,k} = 1, W_{y,l} = 1)P(W_{x,k} = 0, W_{y,l} = 0)}{P(W_{x,k} = 0, W_{y,l} = 1)P(W_{x,k} = 1, W_{y,l} = 0)}, \\
& \quad k \neq l = 1, \ldots, K \text{ or } x \neq y = 1, \ldots, J.
\end{aligned}$$

Table 2.6: GMH and GEE Results of the simulation study for independent strata with $n = 20000$ and $B = 400$.

	mean		$\text{Var}(L/\hat{\gamma})^1_{12}, \text{Var}(L/\hat{\gamma})^2_{12}, \text{Cov}((L/\hat{\gamma})^1_{12}, (L/\hat{\gamma})^2_{12})$		
	$(L^1_{12}, L^1_{12})_{GMH}$	$^{10\cdot}\text{mse}_{GMH}$	$^{10\cdot}\text{emp}_{GMH}$	$^{10\cdot}\text{formulae}_{GMH}$	$^{10\cdot}\text{BT}_{GMH}$
K, N_k, Ψ, θ	$(\hat{\gamma}^1_{12}, \hat{\gamma}^2_{12})_{GEE}$	$^{10\cdot}\text{mse}_{GEE}$	$^{10\cdot}\text{emp}_{GEE}$	$^{10\cdot}\text{robust}_{GEE}$	$^{10\cdot}\text{naive}_{GEE}$
$2, 50, 1, 2$	$-0.000, 0.002$	$1.75, 1.70$	$1.75, 1.70, 0.310$	$1.67, 1.67, 0.279$	$1.77, 1.77, 0.300$
$^{(4)}$	$-0.000, 0.002$	$1.78, 1.73$	$1.78, 1.73, 0.317$	$1.71, 1.71, 0.291$	$1.71, 1.71, 0.291$
$2, 50, 1, 4$	$-0.001, 0.002$	$1.75, 1.71$	$1.75, 1.71, 0.588$	$1.67, 1.67, 0.543$	$1.77, 1.77, 0.584$
$^{(3)}$	$-0.001, 0.002$	$1.78, 1.75$	$1.78, 1.75, 0.601$	$1.71, 1.71, 0.566$	$1.71, 1.71, 0.566$
$^{(2)}2, 50, 4, 2$	$1.425, 1.428$	$2.33, 2.32$	$2.32, 2.30, 0.296$	$2.23, 2.23, 0.294$	$2.55, 2.55, 0.331$
$^{(2)}$	$1.440, 1.443$	$2.39, 2.38$	$2.36, 2.34, 0.304$	$2.27, 2.27, 0.308$	$2.27, 2.27, 0.323$
$^{(1)}2, 50, 4, 4$	$1.429, 1.434$	$2.33, 2.36$	$2.32, 2.34, 0.654$	$2.24, 2.24, 0.611$	$2.56, 2.56, 0.687$
$^{(1)}$	$1.443, 1.450$	$2.39, 2.43$	$2.36, 2.39, 0.668$	$2.27, 2.28, 0.638$	$2.28, 2.28, 0.660$
$2, 100, 1, 2$	$-0.002, -0.002$	$0.84, 0.83$	$0.84, 0.83, 0.148$	$0.82, 0.82, 0.138$	$0.84, 0.84, 0.142$
$^{(3)}$	$-0.002, -0.002$	$0.85, 0.84$	$0.85, 0.84, 0.149$	$0.83, 0.83, 0.141$	$0.83, 0.83, 0.141$
$2, 100, 1, 4$	$-0.002, -0.003$	$0.84, 0.83$	$0.84, 0.83, 0.280$	$0.82, 0.82, 0.269$	$0.84, 0.84, 0.278$
$^{(2)}$	$-0.002, -0.003$	$0.85, 0.84$	$0.85, 0.84, 0.283$	$0.83, 0.83, 0.274$	$0.83, 0.83, 0.275$
$2, 100, 4, 2$	$1.408, 1.405$	$1.08, 1.09$	$1.07, 1.09, 0.144$	$1.07, 1.06, 0.148$	$1.13, 1.13, 0.156$
	$1.415, 1.412$	$1.09, 1.10$	$1.09, 1.10, 0.146$	$1.07, 1.07, 0.151$	$1.08, 1.07, 0.157$
$2, 100, 4, 4$	$1.407, 1.405$	$1.08, 1.09$	$1.08, 1.08, 0.309$	$1.06, 1.06, 0.301$	$1.13, 1.13, 0.317$
	$1.414, 1.412$	$1.10, 1.10$	$1.09, 1.09, 0.313$	$1.07, 1.07, 0.307$	$1.07, 1.07, 0.316$
$20, 5, 1, 2$	$-0.002, 0.005$	$2.25, 2.22$	$2.25, 2.22, 0.382$	$2.12, 2.12, 0.351$	$2.30, 2.30, 0.362$
$^{(17653)}$	$-0.007, -0.005$	$2.83, 2.89$	$2.83, 2.89, 0.527$	$2.47, 2.48, 0.371$	$2.45, 2.45, 0.358$
$20, 5, 1, 4$	$0.003, -0.003$	$2.21, 2.20$	$2.21, 2.20, 0.739$	$2.12, 2.12, 0.678$	$2.29, 2.29, 0.709$
$^{(18027)}$	$0.003, 0.014$	$3.19, 3.00$	$3.19, 3.00, 1.150$	$2.46, 2.46, 0.838$	$2.45, 2.45, 0.824$
$^{(47)}20, 5, 4, 2$	$1.464, 1.460$	$3.56, 3.48$	$3.50, 3.43, 0.391$	$3.22, 3.20, 0.371$	$3.50, 3.50, 0.354$
$^{(19751)}$	$1.890, 1.779$	$7.95, 6.75$	$5.44, 5.22, 0.684$	$3.65, 3.48, 0.319$	$3.55, 3.40, 0.379$
$^{(28)}20, 5, 4, 4$	$1.457, 1.463$	$3.47, 3.47$	$3.42, 3.41, 0.881$	$3.19, 3.20, 0.769$	$3.48, 3.48, 0.743$
$^{(19784)}$	$1.916, 1.945$	$7.78, 9.16$	$4.99, 6.07, 1.565$	$3.68, 3.75, 1.057$	$3.63, 3.69, 1.110$
$20, 10, 1, 2$	$-0.000, -0.003$	$0.93, 0.92$	$0.93, 0.92, 0.162$	$0.91, 0.90, 0.152$	$0.88, 0.88, 0.148$
$^{(1116)}$	$0.000, -0.004$	$1.13, 1.11$	$1.13, 1.11, 0.202$	$1.00, 1.00, 0.170$	$1.00, 1.00, 0.168$
$20, 10, 1, 4$	$-0.001, 0.001$	$0.92, 0.93$	$0.92, 0.93, 0.313$	$0.90, 0.91, 0.298$	$0.88, 0.88, 0.290$
$^{(1227)}$	$-0.002, 0.000$	$1.12, 1.12$	$1.12, 1.12, 0.386$	$1.00, 1.00, 0.334$	$1.00, 1.00, 0.331$
$20, 10, 4, 2$	$1.411, 1.412$	$1.27, 1.28$	$1.27, 1.28, 0.170$	$1.23, 1.24, 0.162$	$1.38, 1.39, 0.167$
$^{(7085)}$	$1.569, 1.570$	$1.90, 1.87$	$1.56, 1.54, 0.219$	$1.35, 1.34, 0.181$	$1.34, 1.34, 0.184$
$20, 10, 4, 4$	$1.412, 1.414$	$1.28, 1.28$	$1.27, 1.27, 0.346$	$1.23, 1.24, 0.335$	$1.39, 1.39, 0.349$
$^{(7667)}$	$1.571, 1.572$	$1.86, 1.91$	$1.52, 1.56, 0.428$	$1.34, 1.34, 0.382$	$1.34, 1.34, 0.383$

$\log(1) = 0, \log(4) = 1.3863$

The value in parentheses is the number of datasets having the sampling zero problem (which are not included)

For a special case of $k = l$, $\Gamma_{xy,kl} = \theta_{xy}$ describes the dependency between items for a given rater k. In contrast, $\Gamma_{xy,kl}$ with $k \neq l$ denotes the dependency between raters, and $x \neq y$ between items. For convenience, we set $\Lambda_{kl} = \Lambda$ for all $k < l = 1, \ldots, K$; $\Gamma_{12,kk} = \theta$ for all $k = 1, \ldots, K$; and $\Gamma_{xy,kl} = \Gamma$ for all $k < l = 1, \ldots, K$ and $x \leq y = 1, 2$.

We first fix the marginal overall rating probabilities $P(Z_k = 1) = 0.5, k = 1, \ldots, K$ and compute the overall rating joint probabilities $P_{\mathbf{Z}} = \{P(Z_1 = z_1, \ldots, Z_K = z_K), z_k = 0, 1; k = 1, \ldots, K\}$ from $\{P(Z_k = 1)\}$ and Λ applying Gange's (1995) method. As in case 1, $\pi_{x|1k}$ is set to be 0.5 for all items $x = 1, \ldots, J$ and strata $k = 1, \ldots, K$. The marginal probabilities $\{\pi_{x|2k}\}$ are computed from the given common odds ratio Ψ. Given a specific overall rating configuration $\mathbf{z} = (z_1, \ldots, z_K)$, the joint distribution $P_{\mathbf{W}|\mathbf{z}} = \{P(W_{1,1} = w_{1,1}, \ldots, W_{J,1} = w_{J,1}, \ldots, W_{1,K} = w_{1,K}, \ldots, W_{J,K} = w_{J,K}|\mathbf{z}), w_{x,k} = 0, 1; x = 1, \ldots, J; k = 1, \ldots, K\}$ can be computed from $\{\pi_{x|ak}\}$, θ and Γ using Gange's method. The 2^K possible overall ratings configurations result in 2^K different joint distributions $P_{\mathbf{W}|\mathbf{z}}$, which are all computed in advance.

Then we draw $N_k = N$ samples from the overall rating joint distribution $P_{\mathbf{Z}}$. Now, given such a realisation z, we can sample one vector of length $J \cdot K$ from $P_{\mathbf{W}|\mathbf{z}}$. Then we separate each of the vectors of length JK into K vectors of length J, such that the k^{th} vector of length J represents the items of rater k. For instance, for $J = 2$, if the k^{th} vector is (0, 1), then it says that rater k selects item 2, but not item 1. We draw samples from $P_{\mathbf{Z}}$ in order to incorporate some dependency in the overall rating between raters.

In case 2, it is not feasible to sample sparse data with a large number of strata ($K >> 5$). Choosing $K = 5$ and $J = 2$, we already get $2^{JK} = 2^{10} = 1024$ joint probabilities in $P_{\mathbf{W}|\mathbf{z}}$ for each overall ratings configuration z. Increasing J or K

creates a problem with a huge number of joint probabilities which is infeasible for most computers. In total, we simulate $n = 20000$ datasets under a variety of configurations. For the bootstrap method, we use the number of bootstrap resamples as $B = 400$. Table 2.7 presents the results using the same notation as in Table 2.6.

Results Table 2.6 shows that the MH approach performs better than the GEE approach, especially when N_k is small. Also, GEE often fails to converge for extremely sparse data, e.g., $N_k = 5$. The convergence problem occurs when the number of parameters increases with the number of strata. In contrast, Table 2.7 shows that GEE provides better estimates for high dependence ($\Gamma \geq 4$) between strata, whereas for low dependence ($\Gamma = 2$) the MH approach still performs as well as GEE.

When we compare the bootstrap with the formulae (co)variances, we can say the following: Under independence of strata the formulae (co)variance and bootstrap (co)variance behave similarly. For the dependent strata case, the bootstrap (co)variance is better than the formulae (co)variance. Only for a few configurations ($\Gamma = 2$) the formulae (co)variance is still quite good and similar to the bootstrap (co)variance despite the violation of the naive independence assumption.

Comparing the (co)variance estimates for GEE, we see that the robust (co)-variance is generally better than the naive as expected, because the naive (co)-variance assumes that the correlation structure chosen is the correct one. In case 1, the dependence only occurs across 2 different items. Since 2 items only require 1 correlation parameter and the choice of working correlation structure "exchangeable" is then automatically correct, the naive and robust (co)variances perform

Table 2.7: GMH and GEE Results of the simulation study for dependent strata with $\Psi = 4$ ($\log(4) = 1.3863$), $\theta = 4$, $n = 20000$ and $B = 400$

	mean		$\mathrm{Var}(L/\hat{\gamma})^1_{12}, \mathrm{Var}(L/\hat{\gamma})^2_{12}, \mathrm{Cov}\{(L/\hat{\gamma})^2_{12}, (L/\hat{\gamma})^2_{12}\}$		
	$(L^1_{12}, L^1_{12})_{GMH}$	$^{10\cdot}\mathrm{mse}_{GMH}$	$^{10\cdot}\mathrm{emp}_{GMH}$	$^{10\cdot}\mathrm{formulae}_{GMH}$	$^{10\cdot}\mathrm{BT}_{GMH}$
K, N_k, Λ, Γ	$(\hat{\gamma}^1_{12}, \hat{\gamma}^2_{12})_{GEE}$	$^{10\cdot}\mathrm{mse}_{GEE}$	$^{10\cdot}\mathrm{emp}_{GEE}$	$^{10\cdot}\mathrm{robust}_{GEE}$	$^{10\cdot}\mathrm{naive}_{GEE}$
$^{(1)}2, 50, 2, 2$	$1.434, 1.431$	$2.41, 2.40$	$2.39, 2.38, 0.728$	$2.24, 2.24, 0.607$	$2.64, 2.63, 0.749$
$^{(1)}$	$1.446, 1.442$	$2.44, 2.43$	$2.40, 2.40, 0.717$	$2.27, 2.27, 0.628$	$2.23, 2.22, 0.358$
$^{(4)}2, 50, 2, 4$	$1.427, 1.432$	$2.57, 2.55$	$2.56, 2.53, 0.830$	$2.24, 2.25, 0.605$	$2.73, 2.74, 0.835$
$^{(4)}$	$1.439, 1.445$	$2.36, 2.37$	$2.33, 2.33, 0.600$	$2.15, 2.16, 0.521$	$2.11, 2.12, 0.454$
$^{(4)}2, 50, 4, 2$	$1.434, 1.426$	$2.48, 2.46$	$2.46, 2.44, 0.779$	$2.24, 2.24, 0.607$	$2.70, 2.68, 0.802$
$^{(4)}$	$1.448, 1.440$	$2.53, 2.48$	$2.49, 2.45, 0.779$	$2.33, 2.32, 0.681$	$2.30, 2.29, 0.432$
$^{(2)}2, 50, 4, 9$	$1.443, 1.443$	$2.86, 2.94$	$2.83, 2.90, 1.148$	$2.28, 2.28, 0.591$	$3.04, 3.05, 1.105$
$^{(2)}$	$1.449, 1.448$	$2.34, 2.36$	$2.30, 2.32, 0.595$	$2.12, 2.12, 0.469$	$2.09, 2.09, 0.640$
$2, 100, 2, 2$	$1.411, 1.411$	$1.13, 1.13$	$1.13, 1.12, 0.347$	$1.07, 1.07, 0.301$	$1.17, 1.17, 0.349$
	$1.417, 1.418$	$1.12, 1.13$	$1.12, 1.12, 0.334$	$1.08, 1.08, 0.310$	$1.05, 1.05, 0.174$
$2, 100, 2, 9$	$1.411, 1.409$	$1.22, 1.25$	$1.22, 1.25, 0.451$	$1.07, 1.07, 0.298$	$1.28, 1.28, 0.446$
	$1.414, 1.412$	$0.96, 0.98$	$0.96, 0.97, 0.183$	$0.93, 0.93, 0.164$	$0.92, 0.92, 0.243$
$2, 100, 4, 2$	$1.408, 1.410$	$1.15, 1.13$	$1.15, 1.13, 0.359$	$1.07, 1.07, 0.301$	$1.19, 1.19, 0.373$
	$1.415, 1.417$	$1.15, 1.15$	$1.14, 1.14, 0.354$	$1.10, 1.10, 0.336$	$1.08, 1.08, 0.210$
$2, 100, 4, 4$	$1.411, 1.407$	$1.21, 1.20$	$1.21, 1.19, 0.429$	$1.07, 1.07, 0.298$	$1.26, 1.26, 0.435$
	$1.416, 1.412$	$1.12, 1.09$	$1.11, 1.09, 0.320$	$1.07, 1.07, 0.299$	$1.04, 1.04, 0.269$
$^{(5)}5, 20, 2, 2$	$1.444, 1.441$	$2.88, 2.79$	$2.84, 2.76, 0.990$	$2.40, 2.39, 0.620$	$3.18, 3.19, 0.978$
$^{(34)}$	$1.500, 1.499$	$2.91, 2.94$	$2.78, 2.81, 0.830$	$2.46, 2.47, 0.663$	$2.50, 2.51, 0.302$
$^{(32)}5, 20, 2, 9$	$1.458, 1.462$	$3.93, 3.83$	$3.87, 3.78, 1.852$	$2.50, 2.50, 0.565$	$4.02, 4.03, 1.705$
$^{(66)}$	$1.497, 1.500$	$2.78, 2.69$	$2.66, 2.56, 0.628$	$2.30, 2.30, 0.488$	$2.02, 2.02, 0.394$
$^{(8)}5, 20, 4, 2$	$1.442, 1.445$	$3.07, 3.12$	$3.03, 3.09, 1.203$	$2.40, 2.41, 0.617$	$3.37, 3.40, 1.179$
$^{(41)}$	$1.500, 1.503$	$3.09, 3.14$	$2.96, 3.01, 1.016$	$2.59, 2.60, 0.805$	$2.67, 2.67, 0.454$
$^{(48)}5, 20, 4, 9$	$1.467, 1.463$	$4.55, 4.42$	$4.49, 4.36, 2.447$	$2.55, 2.53, 0.549$	$4.67, 4.67, 2.317$
$^{(91)}$	$1.500, 1.497$	$2.99, 2.93$	$2.86, 2.80, 0.875$	$2.53, 2.52, 0.701$	$2.22, 2.19, 0.596$
$5, 100, 2, 2$	$1.396, 1.392$	$0.48, 0.47$	$0.48, 0.47, 0.175$	$0.42, 0.42, 0.120$	$0.48, 0.48, 0.169$
	$1.407, 1.404$	$0.44, 0.44$	$0.44, 0.43, 0.131$	$0.43, 0.43, 0.124$	$0.41, 0.41, 0.054$
$5, 100, 2, 4$	$1.397, 1.393$	$0.54, 0.54$	$0.54, 0.53, 0.233$	$0.42, 0.42, 0.119$	$0.54, 0.54, 0.231$
	$1.407, 1.403$	$0.42, 0.41$	$0.41, 0.41, 0.103$	$0.40, 0.40, 0.101$	$0.37, 0.37, 0.063$
$5, 100, 4, 2$	$1.396, 1.391$	$0.52, 0.51$	$0.52, 0.51, 0.213$	$0.42, 0.42, 0.119$	$0.51, 0.51, 0.205$
	$1.407, 1.402$	$0.47, 0.46$	$0.47, 0.46, 0.160$	$0.45, 0.45, 0.150$	$0.44, 0.44, 0.081$
$5, 100, 4, 4$	$1.396, 1.397$	$0.61, 0.62$	$0.61, 0.61, 0.305$	$0.42, 0.42, 0.118$	$0.62, 0.62, 0.305$
	$1.405, 1.406$	$0.45, 0.46$	$0.45, 0.46, 0.143$	$0.44, 0.44, 0.137$	$0.40, 0.40, 0.097$

The value in parentheses is the number of datasets having the sampling zero problem (which are not included)

quite similarly. In case 2, dependence occurs across different items and strata. The performance of the naive (co)variance becomes poor, simply because the "exchangeable" structure now deviates severely from the actual correlation structure of the simulated data.

Most software, like R, only offer simple choices such as "exchangeable", "unstructured", "independence" for all observations and ratings, and one cannot match the exact correlation structure as in our simulation study. The "exchangeable" structure is the most common one, because it incorporates fewer parameters which results in fewer convergence problems.

2.6 Influence Measure

Like the ordinary MH method, one uses the generalised MH estimators when the conditional association between row and column variables remains the same across strata. For multiple responses with J items, we might consider whether homogeneity holds for each of the items simultaneously. One possible way to check the homogeneity uses model fitting. Agresti and Liu (2001) proposed fitting a logit model assuming homogeneity and then to test the goodness-of-fit for the model. It is plausible only when the strata are independent.

However, when the heterogeneity is not severe, the generalised MH estimators still provide a useful descriptive summary of the conditional associations. It might be still important to find out which stratum is "different" from the others. We apply a diagnostics strategy given by Liu and Wang (2007) to the multiple response data. Their influence measure has a similar form to the Cook's distance (Cook 1977)

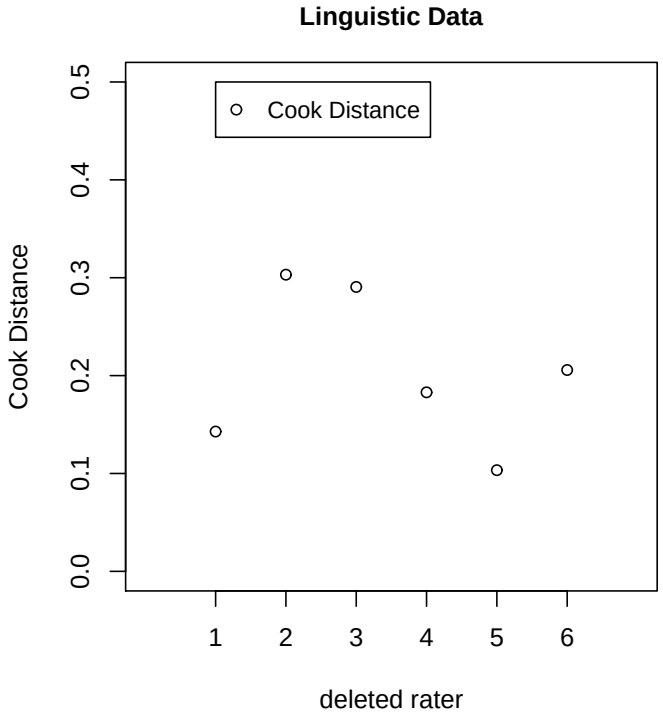
$$CD(\boldsymbol{\beta})_{[d]} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[d]})^T \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[d]})}{p}, \tag{2.17}$$

where $\hat{\boldsymbol{\beta}}$ is a vector of parameter estimates based on all data points; $\boldsymbol{\beta}_{[d]}$ is a vector of parameter estimates with the set d of observations deleted; $\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})$ is the estimate of the covariance matrix for $\hat{\boldsymbol{\beta}}$; and p is the dimension of $\boldsymbol{\beta}$. This influence measure evaluates the difference in estimates due to the deletion. Relatively large values of $CD(\boldsymbol{\beta})_{[d]}$ indicate a high influence of the set d of deleted observations and indicate that they may not follow the given model and/or yield false estimates.

For the generalised MH estimators (2.4), we use (2.17) to find the detail of the heterogeneity across strata for all J items. In $\boldsymbol{\beta}$, we only need to include $J(r-1)$ non-redundant parameters. The others are a linear combination of these $J(r-1)$ parameters, because of the property of the odds ratios where $\bar{L}^x_{ab} + \bar{L}^x_{bc} = \bar{L}^x_{ac}$ for all $a, b, c = 1, \ldots, r$ and $x = 1, \ldots, J$. We let $\hat{\boldsymbol{\beta}} = (\bar{\mathbf{L}}^T_{12}, \ldots, \bar{\mathbf{L}}^T_{1r})^T$, where $\bar{\mathbf{L}}_{ab} = (\bar{L}^1_{ab}, \ldots, \bar{L}^J_{ab})^T$. Let $\hat{\boldsymbol{\beta}}_{[d]}$ be the generalised MH estimates when the d^{th} stratum is deleted. The covariance $\text{Cov}(\hat{\boldsymbol{\beta}})$ is obtained by the bootstrap method. Subsection 2.8.2 on page 88 shows that any set of $J(r-1)$ non-redundant parameters results in the same value of $CD(\boldsymbol{\beta})_{[d]}$.

To determine the heterogeneity between raters in the Linguistics example, Figure 2.1 shows that the second and the third raters have relatively high values of influence measure. This might suggest that the association between rating and items for these two raters differ from the others. When the study is interested in the differences among raters, the influence measure provides a basis for further investigation. The UTI example is not applicable because there are only two strata.

Figure 2.1: Influence Measure for the Linguistics example with single strata deleted



2.7 Conclusion

In this chapter, we use both the model-based (GEE) and non-model-based (MH) approach to evaluate the conditional associations between row and column variables for each of the items for stratified multiple responses. The model-based approach is suitable if one is interested in the model selection in order to find the relationship between the item responses and explanatory variables. For highly sparse data (K large, but N_k small), it might result in convergence problems. However, if one is particularly interested in the conditional association between the item and the explanatory variable given the strata, the MH-type estimators evaluate the association directly. From the simulation studies, the model-based

and non-model-based approach agree with each other.

We give two examples in this paper. The UTI example has independent strata and the linguistic example has dependent strata. For the MH approach with independent strata, Greenland (1989) provided dually consistent variance and covariance estimators for single items, whereas we derived dually consistent covariance estimators between items. For dependent data the bootstrap method provides an easy and plausible way to estimate variances and covariances. It also performs similarly well as the formulae estimates for the independent strata cases.

The deletion influence measure provides a way to evaluate the heterogeneity across strata. Even though it cannot be used directly to test whether the homogeneity assumption holds, it gives a rough idea of the level of heterogeneity.

The proposed MH methods are non-model-based, because we use the Mantel-Haenszel type method. It gives a clear description of the relationship when one focuses on evaluating the conditional association between two variables for each of the items. In general, if the multiple response data has many explanatory variables (> 3), it is more appropriate to describe the relationship among all of them using a model as proposed by Agresti and Liu (2001). However, their models are applicable only for the cases with independent strata.

The linguistic example is a case of multilevel data where there is a hierarchical correlated structure to the data. The responses are correlated within each of the J items; and within each item, the responses are correlated within each of the K raters. Besides the GEE and MH methods, a generalised linear mixed model can also be used for analysing the multilevel data. Fitzmaurice, Laird and Ware (2004) discuss the multilevel generalised linear mixed model. Unfortunately, using the existing software it is not easy to implement the multilevel generalised linear mixed model. Users need to write their own programs for this.

Section 6.5 discusses in detail generalised linear mixed model (GLMM) and describes several algorithms to fit such a model. This gives the reader some impression of the implementation issues he might face when considering a multilevel GLMM. Effects obtained from a GLMM tend to be larger in absolute value than the effects from a GLM or GEE, but so do the standard errors. Therefore messages regarding significance are similar (Agresti and Liu 2001). We expect results from such a multilevel GLMM to be as useful and applicable as the results from the GEE method and the MH approach are, which are considered only in this thesis.

2.8 Proofs

This section provides the missing proofs of the previous sections. In Subsection 2.8.1, we prove the dual consistency of the covariance estimators defined by (2.8), (2.9) and (2.12) on page 51. Then in Subsection 2.8.2, we show a proof for the choice of the influence measure defined in Section 2.6 on page 68.

2.8.1 Proof Covariance Estimators

Preliminaries

Let Var^a, Cov^a and $\mathbb{E}^a$ denote the asymptotic variances, covariances and expectations, whereas Var, Cov and $\mathbb{E}$ are the standard variances, covariances and expectations. For convenience we define $X_A := X^{10}_{xy|ak}$, $X_B := X^{01}_{xy|ak}$, $X_C := X^{11}_{xy|ak}$, $X_D := X^{00}_{xy|ak}$ to avoid confusion with the indices $s, t \in \{0, 1\}$, similarly the π_{st}'s.The number of positive and negative responses were defined as $X_{x|ak}$ and $\bar{X}_{x|ak}$, similarly the probabilities.

We will often suppress subscripts a and k for convenience. We can express π_A, π_B and π_D in terms of $\pi_{x|ak}$, $\pi_{y|ak}$ and π_C

$$\begin{aligned}\pi_A &= \pi_x - \pi_C \\ \pi_B &= \pi_y - \pi_C \\ \pi_D &= 1 - \pi_A - \pi_B - \pi_C = 1 - \pi_x - \pi_y + \pi_C,\end{aligned} \tag{2.18}$$

similarly the $X's$ but replacing 1 with n in the last line of equation (2.18).

Also let $n'_{ak} := n_{ak} - 1$. We assume $\mathbf{X}_{ak} = (X^{00}_{xy|ak}, X^{01}_{xy|ak}, X^{10}_{xy|ak}, X^{11}_{xy|ak})$ follows a multinomial distribution with parameters n_{ak} and $\boldsymbol{\pi}_{ak} = (\pi^{00}_{xy|ak}, \pi^{01}_{xy|ak}, \pi^{10}_{xy|ak}, \pi^{11}_{xy|ak})$; hence, $\mathbb{E}X^2 = nn'\pi^2 + n\pi$ and $\mathbb{E}X_xX_y = nn'\pi_x\pi_y$ $(x \neq y)$.
We compute

$$\begin{aligned}\mathbb{E}X_xX_y &= \mathbb{E}(X_A + X_C)(X_B + X_C) = \mathbb{E}X_C^2 + \mathbb{E}X_AX_B + \mathbb{E}X_AX_C + \mathbb{E}X_BX_C \\ &= nn'\pi_C^2 + n\pi_C + nn'\pi_A\pi_B + nn'\pi_A\pi_C + nn'\pi_B\pi_C \\ &= nn'(\pi_C^2 + \pi_A\pi_B + \pi_A\pi_C + \pi_B\pi_C) + n\pi_C \\ &= nn'(\pi_A + \pi_C)(\pi_B + \pi_C) + n\pi_C \\ &= nn'\pi_x\pi_y + n\pi_C.\end{aligned} \tag{2.19}$$

Using this we find

$$\begin{aligned}\mathbb{E}X_x\bar{X}_y &= \mathbb{E}X_x(n - X_y) \\ &= n\mathbb{E}X_x - \mathbb{E}X_xX_y = n^2\pi_x - nn'\pi_x\pi_y - n\pi_C \\ &= nn'\pi_x\bar{\pi}_y + n(\pi_x - \pi_C) \\ &= nn'\pi_x\bar{\pi}_y + n\pi_A.\end{aligned}$$

Similarly for $\mathbb{E}\bar{X}_x X_y$:

$$\begin{aligned}
\mathbb{E}\bar{X}_x\bar{X}_y &= \mathbb{E}(n - X_x)(n - X_y) \\
&= n^2 - n\mathbb{E}X_x - n\mathbb{E}X_y + \mathbb{E}X_x(n - X_y) \\
&= n^2 - n^2\pi_x - n^2\pi_y + nn'\pi_x\pi_y + n\pi_C \\
&= nn'(1 - \pi_x - \pi_y + \pi_x\pi_y) + n(1 - \pi_x - \pi_y + \pi_C) \\
&= nn'\bar{\pi}_x\bar{\pi}_y + n\pi_D.
\end{aligned}$$

We summarise

$$\begin{aligned}
\mathbb{E}X_x X_y &= nn'\pi_x\pi_y + n\pi_{xy}^{11} \\
\mathbb{E}X_x\bar{X}_y &= nn'\pi_x\bar{\pi}_y + n\pi_{xy}^{10} \\
\mathbb{E}\bar{X}_x X_y &= nn'\bar{\pi}_x\pi_y + n\pi_{xy}^{01} \\
\mathbb{E}\bar{X}_x\bar{X}_y &= nn'\bar{\pi}_x\bar{\pi}_y + n\pi_{xy}^{00}.
\end{aligned} \tag{2.20}$$

Next, we list some useful theorems.

Theorem 2.8.1 (Slutsky's Theorem)**.** *Let $\{X_n, n \geq 1\}$ and $\{Y_n, n \geq 1\}$ be random variables on a probability space. Suppose that $X_n \to_d X$ and $Y_n \to_d c$, where c is a fixed real number. Then (i) $X_n + Y_n \to_d X + c$, (ii)$X_n \cdot Y_n \to_d X \cdot c$, (iii) $X_n/Y_n \to_d X/c$.*

Theorem 2.8.2 (Chebyshev weak law of large numbers)**.** *Let $\{Y_n, n \geq 1\}$ be a random variable with $\mathbb{E}|Y_n|^2 < \infty$, and let $S_n = \sum_{j=1}^n Y_j$. If $\{c_n, n \geq 1\}$ is a sequence of positive constants satisfying Var$(S_n) = o(c_n^2)$, then $\frac{S_n - E(S_n)}{c_n} \to_p 0$.*

Theorem 2.8.3 (Weak law of large numbers)**.** *Let $\{Y_n, n \geq 1\}$ be a sequence of independent and identically distributed random variables, each having a mean $\mathbb{E}Y_n = \mu$ with $\mathbb{E}|Y_n|^2 < \infty$. Then $\frac{1}{n}\sum_{j=1}^n Y_j$ converges in probability to its mean μ.*

Theorem 2.8.4 (Delta method). *If* $\sqrt{n}(\bar{X} - \boldsymbol{\mu}) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Sigma})$ *and* $g = (g_1, \ldots, g_k)$ *is continuously differentiable in a neighbourhood of* $\boldsymbol{\mu}$*, then* $\sqrt{n}(g(\bar{X}) - g(\boldsymbol{\mu})) \rightarrow_d N(\mathbf{0}, \mathbf{B}^T\boldsymbol{\Sigma}\mathbf{B})$ *where* $\mathbf{B}$ *is the partial derivative matrix evaluated at* $\boldsymbol{\mu}$*.*

Theorem 2.8.5 (Sen and Singer (1993, p.123)). *Let* $\{\mathbf{X}_n\}$ *be sequence of random vectors in* $\mathbb{R}^p$ *with mean vectors* $\boldsymbol{\mu}_n$ *and finite covariance matrices* $\boldsymbol{\Sigma}_n, n \geq 1$*, such that*

$$\max_{1\leq i\leq n} \max_{1\leq j\leq p} \mathbb{E}|X_{ij} - \mu_{ij}|^{2+\delta} < \infty \text{ for some } \delta > 0$$

and

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \boldsymbol{\Sigma}_i$$

exists. Then $n^{-1/2} \sum_{i=1}^{n} (\mathbf{X_i} - \boldsymbol{\mu}_i) \rightarrow_d N_p(\mathbf{0}, \boldsymbol{\Sigma})$*.*

Theorem 2.8.6 (Multivariate Central Limit Theorem (C.L.T.)). *Let* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *with* $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})$ *be independent, identically distributed random vectors with mean* $\mathbb{E}(\mathbf{X}_i) = \boldsymbol{\mu}$ *and Var*$(\mathbf{X}_i) = \boldsymbol{\Sigma}$*. Then* $\sqrt{n}(\bar{X} - \boldsymbol{\mu}) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Sigma})$ *with* $\bar{X} = 1/n \sum_{i=1}^{n} \mathbf{X}_i$*.*

Theorem 2.8.7 (Shao (1999, p.39, Theorem 1.8 (vii))). *Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots$ *be random* k*-vectors. Suppose that* $\mathbf{X}_n \rightarrow_d \mathbf{X}$*. Then for any* $r > 0$*,*

$$\lim_{n\to\infty} \mathbb{E}(\|\mathbf{X}_n\|_r)^r = \mathbb{E}(\|\mathbf{X}\|_r)^r < \infty$$

if and only if $\{(\|\mathbf{X}_n\|_r)^r\}$ *is uniformly integrable in the sense that*

$$\lim_{t\to\infty} \sup_n \mathbb{E}[(\|\mathbf{X}_n\|_r)^r \mathbb{1}_{\{\|\mathbf{X}_n\|>t\}}] = 0,$$

where $\|\mathbf{a}\|_p$ *denotes the usual p-norm of vector* $\mathbf{a} = (a_1, \ldots, a_k)$*, for example* $\|\mathbf{a}\|_2 = (\mathbf{a}^T\mathbf{a})^{1/2} = (\sum_{i=1}^{k} a_i^2)^{1/2}$*. Function* $\mathbb{1}_{\{exp\}}$ *is the indicator function and is one if expression* exp *is true and zero otherwise.*

Theorem 2.8.8 (Vaart and Wellner (1996, p.69)). *Let $f : \mathbb{D} \to \mathbb{R}$ be a continuous at every point in set $\mathbb{D}_0 \subset \mathbb{D}$, where $\mathbb{D}$ is a metric space. Let $X_n \to_d X$, where X takes its values in $\mathbb{D}_0$. If $f(X_n)$ is uniformly integrable, then $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$.*

Derivation of Asymptotic Covariances and Variances

Let $N = \sum_k N_k$ and as $N \to \infty$ let $N\alpha_{ak} = n_{ak}$, where $0 < \alpha_{ak} < 1$. Thus $N_k = \sum_i n_{ik} = N \sum_i \alpha_{ik}$. Recall $L^x_{ab} = \log \hat{\Psi}^x_{ab} = \log C_{x|ab}/C_{x|ba}$.
We can write

$$\hat{\Psi}^x_{ab} - \Psi^x_{ab} = \frac{C_{x|ab} - \Psi^x_{ab} C_{x|ba}}{C_{x|ba}} = \frac{(C_{x|ab} - \Psi^x_{ab} C_{x|ba})/M}{C_{x|ba}/M} = \frac{\Omega_{x|ab}/M}{C_{x|ba}/M}. \tag{2.21}$$

with $\omega_{x|abk} := c_{x|abk} - \Psi_{x|ab} c_{x|bak}$ and $\Omega = \sum_k \omega_k$. Notation M can stand for either N or K.

First we consider the asymptotics for $C_{x|ab}$ under both limiting models. The term $c_{x|ab}$ is a bounded random variable under Model II, hence, the variance of $C_{x|ab}$ is $o(K^2)$. We apply the Chebyshev weak law of large numbers and have

$$C_{x|ab}/K = \sum_{k=1}^{K} c_{x|ab}/K \overset{K\to\infty}{\longrightarrow} \lim_{K\to\infty} \sum_{k=1}^{K} \mathbb{E}c_{x|ab}/K = \lim_{K\to\infty} \mathbb{E}C_{x|ab}/K. \tag{2.22}$$

This limit is finite and nonzero. Under model I

$$\begin{aligned} C_{x|ab}/N &= \sum_{k=1}^{K} c_{x|ab}/N = \sum_{k=1}^{K} X_{x|ak}\bar{X}_{x|bk}/(N_k N) \\ &= \sum_{k=1}^{K} \frac{n_{ak}n_{bk}}{N_k N} \frac{X_{x|ak}}{n_{ak}} \frac{\bar{X}_{x|bk}}{n_{bk}} = \sum_{k=1}^{K} \frac{n_{ak}n_{bk}}{NN} \frac{N}{N_k} \frac{X_{x|ak}}{n_a} \frac{\bar{X}_{x|bk}}{n_b} \\ &\overset{N\to\infty}{\longrightarrow} \sum_{k=1}^{K} \alpha_{ak}\alpha_{bk} (\sum_i \alpha_{ik})^{-1} \pi_{x|ak}\bar{\pi}_{x|bk} = \sum_{k=1}^{K} (\sum_i \alpha^{-1}_{ik})^{-1} \pi_{x|ak}\bar{\pi}_{x|bk}, \end{aligned} \tag{2.23}$$

and for the term $\mathbb{E}C_{x|ab}/N$ we derive

$$\begin{aligned}\mathbb{E}C_{x|ab}/N &= \sum_{k=1}^{K}\mathbb{E}c_{x|abk}/N = \sum_{k=1}^{K}\mathbb{E}X_{x|ak}\mathbb{E}\bar{X}_{x|bk}/(N_kN)\\ &= \sum_{k=1}^{K}\frac{n_{ak}n_{bk}}{N_kN}\pi_{x|ak}\pi_{x|bk} = \sum_{k=1}^{K}\frac{n_{ak}n_{bk}}{NN}\frac{N}{N_k}\pi_{x|ak}\bar{\pi}_{x|bk}\\ \overset{N\to\infty}{\longrightarrow} &\sum_{k=1}^{K}\alpha_{ak}\alpha_{bk}(\sum_i\alpha_{ik})^{-1}\pi_{x|ak}\pi_{x|bk} = \sum_{k=1}^{K}(\sum_i\alpha_{ik}^{-1})^{-1}\pi_{x|ak}\bar{\pi}_{x|bk}. \qquad (2.24)\end{aligned}$$

The expectation splits into two due to the independence of rows a and b. Note we use the equality $\alpha_{ak}\alpha_{bk}(\sum_i\alpha_{ik})^{-1} = (\sum_i\alpha_{ik}^{-1})^{-1}$ for convenience, although it is only true for $r = 2$. Hence we conclude by comparing (2.23) with (2.24) and taking into account (2.22), that $\mathbb{E}C_{x|ab}/M$ converges under both limiting models to a constant $\lim_{M\to\infty}\mathbb{E}C_{x|ab}/M$

$$\mathbb{E}C_{x|ab}/M \overset{M\to\infty}{\longrightarrow} \lim_{M\to\infty}\mathbb{E}C_{x|ab}/M < \infty \text{ with } M \in \{N, K\}. \qquad (2.25)$$

We also have $\mathbb{E}c_{x|ab} = \Psi_{ab}^x\mathbb{E}c_{x|ba}$ by the common odds ratio assumption (2.2), hence

$$\lim_{M\to\infty}\mathbb{E}C_{x|ab}/M = \Psi_{ab}^x\lim_{M\to\infty}\mathbb{E}C_{x|ba}/M \text{ with } M \in \{N, K\}. \qquad (2.26)$$

Now we can write for both limiting models assuming that the asymptotic covariances exist

$$\begin{aligned}&\lim_{M\to\infty} M\cdot\text{Cov}^a(L_{ab}^x, L_{ac}^y)\\ &= 1/(\Psi_{ab}^x\Psi_{ac}^y)\lim_{M\to\infty} M\cdot\text{Cov}^a(\hat{\Psi}_{ab}^x, \hat{\Psi}_{ac}^y)\\ &= 1/(\Psi_{ab}^x\Psi_{ac}^y)\frac{\lim_{M\to\infty} M\cdot\text{Cov}^a(\Omega_{x|ab}/M, \Omega_{y|ac}/M)}{(\lim_{M\to\infty}\mathbb{E}C_{x|ba}/M)(\lim_{M\to\infty}\mathbb{E}C_{y|ca}/M)}\\ &= \frac{\lim_{M\to\infty} M\cdot\text{Cov}^a(\Omega_{x|ab}/M, \Omega_{y|ac}/M)}{(\lim_{M\to\infty}\mathbb{E}C_{x|ab}/M)(\lim_{M\to\infty}\mathbb{E}C_{y|ac}/M)} \text{ with } M \in \{N, K\}. \qquad (2.27)\end{aligned}$$

The first equality follows from the delta method (Theorem 2.8.4), the second from (2.21), (2.25) and Slutsky's theorem (Theorem 2.8.1), and the final equality follows from (2.26).

"Sparse Data" Limiting Model First by independence of rows $\text{Cov}(\Omega_{x|ab}, \Omega_{y|cd}) = \sum_{k=1}^{K} \text{Cov}(\omega_{x|ab}, \omega_{y|cdk})$. We will use either expression dependent on which is more convenient. Note that $\mathbb{E}|\omega_{x|abk} - \mathbb{E}\omega_{x|abk}|^3 = \mathbb{E}|\omega_{x|abk}|^3 = O(1)$, because $c_{x|abk}$ is a bounded random variable under model II. By setting $\delta = 1$, we conclude from Theorem 2.8.5 that $K^{-1/2}\ (\Omega_{x|ab},\ \Omega_{y|ac}) = \sqrt{K}(\Omega_{x|ab}/K,\ \Omega_{y|ac}\ /K)$ converges to a zero mean multivariate normal distribution with covariance $\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \text{Cov}(\omega_{x|abk}, \omega_{y|ack})$, by noting that $\mathbb{E}\omega_{abk} = 0$ and $\text{Cov}(\omega_{x|ab},\ \omega_{y|ac})$ exists. We can write

$$\lim_{K\to\infty} K \cdot \text{Cov}^a(\Omega_{x|ab}/K, \Omega_{y|ac}/K) = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \text{Cov}(\omega_{x|abk}, \omega_{y|ack}). \tag{2.28}$$

"Large Stratum" Limiting Model Under model I, $\sqrt{N}(\Omega_{x|ab}/N, \Omega_{y|ac}/N)$ converges by the delta method (Theorem 2.8.4) to a zero mean multivariate normal distribution with covariance V, because $\Omega_{x|ab}$ is a function of the sample proportions, which converge by the central limit theorem (C.L.T.) to a normal distribution. The delta method provides an explicit formula for this asymptotic variance V. We want to show now that $V[= \lim_{N\to\infty} N \cdot \text{Cov}^a(\frac{1}{N}\Omega_{x|ab},\ \frac{1}{N}\Omega_{y|ac})]$ is identical to $\lim_{N\to\infty} N \cdot \text{Cov}(\frac{1}{N}\Omega_{x|ab},\ \frac{1}{N}\Omega_{y|ac})$, that is

$$V \equiv \lim_{N\to\infty} N \cdot \text{Cov}(\frac{1}{N}\Omega_{x|ab}, \frac{1}{N}\Omega_{y|ac}) \left[= \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{K} \text{Cov}(\omega_{x|abk}, \omega_{y|ack}) \right]. \tag{2.29}$$

Let X_n and Y_n be a sequence of random variables (r.v.). If X_n^r is uniformly integrable, so is X_n^s with $s < r$, or more generally, if $X_n \leq Y$ and Y is uniformly

integrable, then X_n is also uniformly integrable. We can formulate the following Lemma.

Lemma 1. *Let $(X_n, Y_n) \to_d (X, Y)$ with $\mathbb{E}X_n^2 < \infty$ and $\mathbb{E}Y_n^2 < \infty$. If X_n^2 and Y_n^2 are uniformly integrable, then Cov(X_n, Y_n) converges to Cov(X, Y).*

Proof. Without loss of generality (w.l.o.g.) $\mathbb{E}X_n = \mathbb{E}Y_n = 0$. From $|X_n \cdot Y_n| > t$ follows $|X_n| > \sqrt{t}$ or $|Y_n| > \sqrt{t}$, hence, $\mathbb{1}_{\{|X_n \cdot Y_n| > t\}} \leq \mathbb{1}_{\{|X_n| > \sqrt{t}\}} + \mathbb{1}_{\{|Y_n| > \sqrt{t}\}}$. Define $\tilde{X}_n := |X_n| \cdot \mathbb{1}_{\{|X_n| > \sqrt{t}\}}$ and $\tilde{Y}_n := |Y_n| \cdot \mathbb{1}_{\{|Y_n| > \sqrt{t}\}}$. From the uniform integrability of X_n^2 and Y_n^2 and Theorem 2.8.7 follows that $\sup_n \mathbb{E}\tilde{X}_n$, $\sup_n \mathbb{E}\tilde{Y}_n$, $\sup_n \mathrm{Var}(\tilde{X}_n)$ and $\sup_n \mathrm{Var}(\tilde{Y}_n)$ converge to zero (as t goes to infinity so does $\sqrt{t}$), and that $\sup_n \mathrm{Var}(|X_n|)$, $\sup_n \mathrm{Var}(|Y_n|)$, $\mathbb{E}|X_n|$ and $\mathbb{E}|Y_n|$ are finite.
We have

$$
\begin{aligned}
&\sup_n \mathbb{E}|X_n \cdot Y_n| \cdot \mathbb{1}_{\{|X_n \cdot Y_n| > t\}} \\
\leq &\sup_n \mathbb{E}|X_n \cdot Y_n| \left(\mathbb{1}_{\{|X_n| > \sqrt{t}\}} + \mathbb{1}_{\{|Y_n| > \sqrt{t}\}} \right) = \sup_n \left[\mathbb{E}|X_n| \cdot \tilde{Y}_n + \mathbb{E}\tilde{X}_n \cdot |Y_n| \right] \\
\leq &\sup_n \mathbb{E}|X_n| \cdot \tilde{Y}_n + \sup_n \mathbb{E}\tilde{X}_n \cdot |Y_n| \\
\leq &\sup_n (\mathrm{Var}(|X_n|))^{1/2} \cdot (\mathrm{Var}(\tilde{Y}_n))^{1/2} + \sup_n \mathbb{E}|X_n| \cdot \mathbb{E}\tilde{Y}_n \\
&+ \sup_n (\mathrm{Var}(\tilde{X}_n))^{1/2} \cdot (\mathrm{Var}(|Y_n|))^{1/2} + \sup_n \mathbb{E}\tilde{X}_n \cdot \mathbb{E}|Y_n| \\
\overset{t \to \infty}{\longrightarrow} &\sup_n (\mathrm{Var}(X_n))^{1/2} \cdot 0 + \sup_n \mathbb{E}|X_n| \cdot 0 + 0 \cdot \sup_n (\mathrm{Var}(Y_n))^{1/2} + 0 \cdot \mathbb{E}|Y_n| = 0.
\end{aligned}
$$

From $|\mathbb{E}(|X_n| \cdot \tilde{Y}_n) - (\mathbb{E}|X_n| \cdot \mathbb{E}\tilde{Y}_n)| = |\mathrm{Cov}(|X_n|, \tilde{Y}_n)| \leq \mathrm{Var}(|X_n|)^{1/2} \cdot \mathrm{Var}(\tilde{Y}_n)^{1/2}$ and $\mathbb{E}|X_n| \cdot \mathbb{E}\tilde{Y}_n \geq 0$ follows that

$$
\mathbb{E}(|X_n| \cdot \tilde{Y}_n) = |(\mathbb{E}|X_n| \cdot \tilde{Y}_n)| \leq \mathrm{Var}(|X_n|)^{1/2} \cdot \mathrm{Var}(\tilde{Y}_n)^{1/2} + \mathbb{E}|X_n| \cdot \mathbb{E}\tilde{Y}_n,
$$

that is line 4, similarly line 5. We showed the uniform integrability of $f(X_n, Y_n)$

with $f(a,b) = a \cdot b$. By Theorem 2.8.8 and using that $\mathbb{R}^2$ is a metric space, $\mathbb{E}f(X_n, Y_n) = \mathbb{E}X_nY_n = \text{Cov}(X_n, Y_n)$ converges to $\mathbb{E}f(X,Y) = \mathbb{E}XY = \text{Cov}(X,Y)$.

□

Remark 2.8.9. Given $(X_n, Y_n) \to_d (X, Y)$, it would be very suprising if $\text{Var}(X_n) \to \text{Var}(X)$ and $\text{Var}(Y_n) \to \text{Var}(Y)$ (which follows from the uniform integrability of X_n^2 and Y_n^2 with $\mathbb{E}X_n^2 < \infty$ and $\mathbb{E}Y_n^2 < \infty$) but $\text{Cov}(X_n, Y_n) \nrightarrow \text{Cov}(X,Y)$.

Later in Chapter 3, see Remark 3.3.2 on page 98, we show the uniform integrability of $(\sqrt{N} \cdot \Omega_{x|ab}/N)^2 = (N^{-1/2} \cdot \Omega_{x|ab})^2$. It follows from Lemma 1 that $\text{Cov}(N^{-1/2} \cdot \Omega_{x|ab}, N^{-1/2} \cdot \Omega_{y|ac})$ converges to $\lim_{N\to\infty} \text{Cov}^a(N^{-1/2} \cdot \Omega_{x|ab}, N^{-1/2} \cdot \Omega_{y|ac})$, or equivalently $N \cdot \text{Cov}(\frac{1}{N}\Omega_{x|ab}, \frac{1}{N}\Omega_{y|ac})$ converges to $V = \lim_{N\to\infty} N \cdot \text{Cov}^a(\frac{1}{N}\Omega_{x|ab}, \frac{1}{N}\Omega_{y|ac})$.

Asymptotic Covariance for Both Limiting Models We express the asymptotic covariances for both limiting models expressed in equation (2.27) by using (2.28) and (2.29) as

$$\lim_{M\to\infty} M \cdot \text{Cov}^a(L_{ab}^x, L_{ac}^y) = \frac{\lim_{M\to\infty} \frac{1}{M} \sum_k \text{Cov}(\omega_{x|abk}, \omega_{y|ack})}{(\lim_{M\to\infty} \frac{1}{M}\mathbb{E}C_{x|ab})(\lim_{M\to\infty} \frac{1}{M}\mathbb{E}C_{y|ac})} \tag{2.30}$$

for $M \in \{K, N\}$.

Remark 2.8.10. If we want to compute the asymptotic variance for model I, we can apply the delta method. However, we think these computations are more costly than computing simply the limit of $\lim_N \frac{1}{N} \sum_k \text{Cov}(\omega_{x|abk}, \omega_{y|ack})$. The computation of $\text{Cov}(\omega_{x|abk}, \omega_{y|ack})$ is not cheap either, but is a by-product of the computation of the "sparse-data" limiting variance.

Computation of Cov$(\omega_{x|abk}, \omega_{y|ack})$ Now we compute $N_k^2\text{Cov}(\omega_{x|ab}, \omega_{y|ab})$ using (2.20):

$$
\begin{aligned}
&N_k^2\text{Cov}(c_{x|ab} - \Psi_{ab}^x c_{x|ba}, c_{y|ab} - \Psi_{ab}^y c_{y|ba}) \\
&= \mathbb{E}(c_{x|ab} - \Psi_{ab}^x c_{x|ba})(c_{y|ab} - \Psi_{ab}^y c_{y|ba}) - \mathbb{E}(c_{x|ab} - \Psi_{ab}^x c_{x|ba})\mathbb{E}(c_{y|ab} - \Psi_{ab}^y c_{y|ba}) \\
&= \mathbb{E}(c_{x|ab} - \Psi_{ab}^x c_{x|ba})(c_{y|ab} - \Psi_{ab}^y c_{y|ba}) \\
&= \mathbb{E}c_{x|ab}c_{y|ab} - \Psi_{ab}^x c_{x|ba}c_{y|ab} - \Psi_{ab}^y \mathbb{E}c_{x|ab}c_{y|ba} + \Psi_{ab}^x\Psi_{ab}^y \mathbb{E}c_{x|ba}c_{y|ba} \\
&= \mathbb{E}X_{x|a}X_{y|a}\mathbb{E}\bar{X}_{x|b}\bar{X}_{y|b} - \Psi_{ab}^x \mathbb{E}\bar{X}_{x|a}X_{y|a}\mathbb{E}X_{x|b}\bar{X}_{y|b} \\
&- \Psi_{ab}^y EX_{x|a}\bar{X}_{y|a}\mathbb{E}\bar{X}_{x|b}X_{y|b} + \Psi_{ab}^x\Psi_{ab}^y \mathbb{E}\bar{X}_{x|a}\bar{X}_{y|a}\mathbb{E}X_{x|b}X_{y|b} \\
&= n_a n_b\{(n_a'\pi_{x|a}\pi_{y|a} + \pi_{xy|a}^{11})(n_b'\bar{\pi}_{x|b}\bar{\pi}_{y|b} + \pi_{xy|b}^{00}) \\
&- \Psi_{ab}^x(n_a'\bar{\pi}_{x|a}\pi_{y|a} + \pi_{xy|a}^{01})(n_b'\pi_{x|b}\bar{\pi}_{y|b} + \pi_{xy|b}^{10}) \\
&- \Psi_{ab}^y(n_a'\pi_{\pi|a}\bar{\pi}_{y|a} + \pi_{xy|a}^{10})(n_b'\bar{\pi}_{x|b}\pi_{y|b} + \pi_{xy|b}^{01}) \\
&+ \Psi_{ab}^x\Psi_{ab}^y(n_a'\bar{\pi}_{x|a}\bar{\pi}_{y|a} + \pi_{xy|a}^{00})(n_b'\pi_{x|b}\pi_{y|b} + \pi_{xy|b}^{11})\} \\
&= n_a n_b\{n_a'n_b'(\pi_{x|a}\pi_{y|a}\bar{\pi}_{x|b}\bar{\pi}_{y|b} - \Psi_{ab}^x\bar{\pi}_{x|a}\pi_{y|a}\pi_{x|b}\bar{\pi}_{y|b} \\
&- \Psi_{ab}^y\pi_{x|a}\bar{\pi}_{y|a}\bar{\pi}_{x|b}\pi_{y|b} + \Psi_{ab}^x\Psi_{ab}^y\bar{\pi}_{x|a}\bar{\pi}_{y|a}\pi_{x|b}\pi_{y|b}) \\
&+ n_a'(\pi_{x|a}\pi_{y|a}\pi_{xy|b}^{00} - \Psi_{ab}^x\bar{\pi}_{x|a}\pi_{y|a}\pi_{xy|b}^{10} - \Psi_{ab}^y\pi_{\pi|a}\bar{\pi}_{y|a}\pi_{xy|b}^{01} + \Psi_{ab}^x\Psi_{ab}^y\bar{\pi}_{x|a}\bar{\pi}_{y|a}\pi_{xy|b}^{11}) \\
&+ n_b'(\pi_{xy|a}^{11}\bar{\pi}_{x|b}\bar{\pi}_{y|b} - \Psi_{ab}^x\pi_{xy|a}^{01}\pi_{x|b}\bar{\pi}_{y|b} - \Psi_{ab}^y\pi_{xy|a}^{10}\bar{\pi}_{x|b}\pi_{y|b} + \Psi_{ab}^x\Psi_{ab}^y\pi_{xy|a}^{00}\pi_{x|b}\pi_{y|b}) \\
&+ (\pi_{xy|a}^{11}\pi_{xy|b}^{00} - \Psi_{ab}^x\pi_{xy|a}^{01}\pi_{xy|b}^{10} - \Psi_{ab}^y\pi_{xy|a}^{10}\pi_{xy|b}^{01} + \Psi_{ab}^x\Psi_{ab}^y\pi_{xy|a}^{00}\pi_{xy|b}^{11})\} \\
&= n_a n_b n_a' n_b'\pi_{x|a}\pi_{y|a}\bar{\pi}_{x|b}\bar{\pi}_{y|b}(+1 - 1 - 1 + 1) \\
&+ n_a n_b\{n_a'(\pi_{x|a}\pi_{y|a}\pi_{xy|b}^{00} - \Psi_{ab}^x\bar{\pi}_{x|a}\pi_{y|a}\pi_{xy|b}^{10} - \Psi_{ab}^y\pi_{\pi|a}\bar{\pi}_{y|a}\pi_{xy|b}^{01} + \Psi_{ab}^x\Psi_{ab}^y\bar{\pi}_{x|a}\bar{\pi}_{y|a}\pi_{xy|b}^{11}) \\
&+ n_b'(\pi_{xy|a}^{11}\bar{\pi}_{x|b}\bar{\pi}_{y|b} - \Psi_{ab}^x\pi_{xy|a}^{01}\pi_{x|b}\bar{\pi}_{y|b} - \Psi_{ab}^y\pi_{xy|a}^{10}\bar{\pi}_{x|b}\pi_{y|b} + \Psi_{ab}^x\Psi_{ab}^y\pi_{xy|a}^{00}\pi_{x|b}\pi_{y|b}) \\
&+ (\pi_{xy|a}^{11}\pi_{xy|b}^{00} - \Psi_{ab}^x\pi_{xy|a}^{01}\pi_{xy|b}^{10} - \Psi_{ab}^y\pi_{xy|a}^{10}\pi_{xy|b}^{01} + \Psi_{ab}^x\Psi_{ab}^y\pi_{xy|a}^{00}\pi_{xy|b}^{11})\} \\
&= N_k^2\{d_{xy|ab}^{11} - \Psi_{ab}^x d_{xy|ab}^{01} - \Psi_{ab}^y d_{xy|ab}^{10} + \Psi_{ab}^x\Psi_{ab}^y d_{xy|ab}^{00}\}
\end{aligned}
$$

with

$$d^{st}_{xy|ab} = \frac{n_a n_b}{N_k^2}\{n'_a \pi^s_{x|a}\pi^t_{y|a}\pi^{\bar{s}\bar{t}}_{xy|b} + n'_b \pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|b}\pi^{st}_{xy|a} + \pi^{\bar{s}\bar{t}}_{xy|a}\pi^{st}_{xy|b}\}, \tag{2.31}$$

where $\pi^1_{x|a} := \pi_{x|a}$ and $\pi^0_{x|a} := \bar{\pi}_{x|a}$ with $\bar{s} := 1 - s$. The second equality follows from $\mathbb{E}(c_{x|ab} - \Psi^x_{ab}c_{x|ba}) = 0$ and the 7th from $\pi_{x|a}\pi_{y|a}\bar{\pi}_{x|b}\bar{\pi} = \Psi^x_{ab}\bar{\pi}_{x|a}\pi_{y|a}\pi_{x|b}\bar{\pi}_{y|b} = \Psi^y_{ab}\pi_{x|a}\bar{\pi}_{y|a}\bar{\pi}_{x|b}\pi_{y|b} = \Psi^x_{ab}\Psi^y_{ab}\bar{\pi}_{x|a}\bar{\pi}_{y|a}\pi_{x|b}\pi_{y|b}$ by the assumption of a common odds ratio.

Similarly we obtain

$$\begin{aligned}
&N_k^2\text{Cov}(c_{x|ab} - \Psi^x_{ab}c_{x|ba}, c_{y|ac} - \Psi^y_{ac}c_{y|ca})\\
&= \mathbb{E}(c_{x|ab} - \Psi^x_{ab}c_{x|ba})(c_{y|ac} - \Psi^y_{ac}c_{y|ca})\\
&= \mathbb{E}c_{x|ab}c_{y|ac} - \Psi^x_{ab}c_{x|ba}c_{y|ac} - \Psi^y_{ac}\mathbb{E}c_{x|ab}c_{y|ca} + \Psi^x_{ab}\Psi^y_{ac}\mathbb{E}c_{x|ca}c_{y|ca}\\
&= \mathbb{E}X_{x|a}X_{y|a}\mathbb{E}\bar{X}_{x|b}\mathbb{E}\bar{X}_{y|c} - \Psi^x_{ab}\mathbb{E}\bar{X}_{x|a}X_{y|a}\mathbb{E}X_{x|b}\mathbb{E}\bar{X}_{y|c}\\
&- \Psi^y_{ab}\mathbb{E}X_{x|a}\bar{X}_{y|a}\mathbb{E}\bar{X}_{x|b}\mathbb{E}X_{y|c} + \Psi^x_{ab}\Psi^y_{ab}\mathbb{E}\bar{X}_{x|a}\bar{X}_{y|a}\mathbb{E}X_{x|b}\mathbb{E}X_{y|c}\\
&= n_a n_b n_c\{(n'_a\pi_{x|a}\pi_{y|a} + \pi^{11}_{xy|a})\bar{\pi}_{x|b}\bar{\pi}_{y|c} - \Psi^x_{ab}(n'_a\bar{\pi}_{x|a}\pi_{y|a} + \pi^{01}_{xy|a})\pi_{x|b}\bar{\pi}_{y|b}\\
&- \Psi^y_{ab}(n'_a\pi_{\pi|a}\bar{\pi}_{y|a} + \pi^{10}_{xy|a})\bar{\pi}_{x|b}\pi_{y|b} + \Psi^x_{ab}\Psi^y_{ab}(n'_a\bar{\pi}_{x|a}\bar{\pi}_{y|a} + \pi^{00}_{xy|a})\pi_{x|b}\pi_{y|b}\}\\
&= n_a n_b n_c\{\pi^{11}_{xy|a}\bar{\pi}_{x|b}\bar{\pi}_{y|c} - \Psi^x_{ab}\pi^{01}_{xy|a}\pi_{x|b}\bar{\pi}_{y|c} - \Psi^y_{ab}\pi^{10}_{xy|a}\bar{\pi}_{x|b}\pi_{y|c} + \Psi^x_{ab}\Psi^y_{ab}\pi^{00}_{xy|a}\pi_{x|b}\pi_{y|c}\}\\
&+ \pi_{x|a}\pi_{y|a}\bar{\pi}_{x|b}\bar{\pi}_{y|c}\{+1 - 1 - 1 + 1\}\\
&= N_k^2\{d^{11}_{abc} - \Psi^x_{ab}d^{01}_{abc} - \Psi^y_{ab}d^{10}_{abc} + \Psi^x_{ab}\Psi^y_{ab}d^{00}_{abc}\}
\end{aligned}$$

with

$$d^{st}_{abc} = \frac{n_a n_b n_c}{N_k^2}\pi^{st}_{xy|a}\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|c}.$$

If indices a, b, c, d are all distinct, $\text{Cov}(\omega_{x|abk}, \omega_{y|cdk}) = 0$ owing to the independence of rows.

Now we write using (2.26) and D representing $\sum_k d$

$$\begin{aligned}
&\lim_{M\to\infty} M \cdot \text{Cov}^a(L^x_{ab}, L^y_{ab}) \\
&= \frac{\lim_{M\to\infty} \frac{1}{M}\{D^{11}_{ab} - \Psi^x_{ab} D^{01}_{ab} - \Psi^y_{ab} D^{10}_{ab} + \Psi^x_{ab}\Psi^y_{ab} D^{00}_{ab}\}}{\lim_{M\to\infty} \frac{1}{M}\mathbb{E}C_{x|ab} \lim_{M\to\infty} \frac{1}{M}\mathbb{E}C_{y|ab}} \\
&= \frac{\lim \frac{1}{M} D^{11}_{ab}}{\lim \frac{1}{M}\mathbb{E}C_{x|ab} \frac{1}{M}\mathbb{E}C_{y|ab}} - \frac{\lim \frac{1}{M} D^{01}_{ab}}{\lim \frac{1}{M}\mathbb{E}C_{x|ba} \frac{1}{M}\mathbb{E}C_{y|ab}} \\
&- \frac{\lim \frac{1}{M} D^{10}_{ab}}{\lim \frac{1}{M}\mathbb{E}C_{x|ab} \frac{1}{M}\mathbb{E}C_{y|ba}} + \frac{\lim \frac{1}{M} D^{00}_{ab}}{\lim \frac{1}{M}\mathbb{E}C_{x|ba} \frac{1}{M}\mathbb{E}C_{y|ba}}.
\end{aligned} \tag{2.32}$$

Similarly,

$$\begin{aligned}
&\lim_{M\to\infty} M \cdot \text{Cov}^a(L^x_{ab}, L^y_{ac}) \\
&= \frac{\lim \frac{1}{M} D^{11}_{abc}}{\lim \frac{1}{M}\mathbb{E}C_{x|abc} \frac{1}{M} \lim \mathbb{E}C_{y|ab}} - \frac{\lim \frac{1}{M} D^{01}_{abc}}{\lim \frac{1}{M}\mathbb{E}C_{x|ba} \lim \frac{1}{M}\mathbb{E}C_{y|ab}} \\
&- \frac{\lim \frac{1}{M} D^{10}_{abc}}{\lim \frac{1}{M}\mathbb{E}C_{x|ab} \frac{1}{M} \lim \mathbb{E}C_{y|ba}} + \frac{\lim \frac{1}{M} D^{00}_{abc}}{\lim \frac{1}{M}\mathbb{E}C_{x|ba} \lim \frac{1}{M}\mathbb{E}C_{y|ba}}.
\end{aligned} \tag{2.33}$$

Dual Consistency of Covariance Estimators

The estimators $U_{xy|abb}$ and $U_{xy|abc}$ for $\text{Cov}(L^x_{ab}, L^y_{ab})$ and $\text{Cov}(L^x_{ab}, L^y_{ac})$, respectively, are defined by (2.8) and (2.9) and the $\hat{d}$'s estimating the d's by (2.10) and (2.11). Next we show that $U_{xy|abb}$ and $U_{xy|abc}$ are dually consistent, hence, we must show that $\lim_{M\to\infty} M \cdot U_{xy|ab} = \lim_{M\to\infty} M \cdot \text{Cov}(L^x_{ab}, L^y_{ab})$ and $\lim_{M\to\infty} U_{xy|abc} = \lim_{M\to\infty} M \cdot \text{Cov}(L^x_{ab}, L^y_{ac})$.

We can write

$$\lim_{M\to\infty} M \cdot U_{xy|abb}$$

$$= \frac{\lim \frac{1}{M}\hat{D}_{ab}}{\lim \frac{1}{M}C_{x|ab} \lim \frac{1}{M}C_{y|ab}} - \frac{\lim \frac{1}{M}\hat{D}^{x}_{ab}}{\lim \frac{1}{M}C_{x|ba} \lim \frac{1}{M}C_{y|ab}}$$
$$- \frac{\lim \frac{1}{M}\hat{D}^{y}_{ab}}{\lim \frac{1}{M}C_{x|ab} \lim \frac{1}{M}C_{y|ba}} + \frac{\lim \frac{1}{M}\hat{D}^{xy}_{ab}}{\lim \frac{1}{M}C_{x|ba} \lim \frac{1}{M}C_{y|ba}}, \tag{2.34}$$

similarly,

$$\lim_{M\to\infty} M \cdot U_{xy|abc}$$
$$= \frac{\lim \frac{1}{M}\hat{D}_{abc}}{\lim \frac{1}{M}C_{x|ab} \lim \frac{1}{M}C_{y|ac}} - \frac{\lim \frac{1}{M}\hat{D}^{x}_{abc}}{\lim \frac{1}{M}C_{x|ba} \lim \frac{1}{M}C_{y|ac}}$$
$$- \frac{\lim \frac{1}{M}\hat{D}^{y}_{abc}}{\lim \frac{1}{M}C_{x|ab} \lim \frac{1}{M}C_{y|ca}} + \frac{\lim \frac{1}{M}\hat{D}^{xy}_{abc}}{\lim \frac{1}{M}C_{x|ba} \lim \frac{1}{M}C_{y|ca}}. \tag{2.35}$$

Comparing (2.34) and (2.35) with (2.32) and (2.33), it remains to show

$$\lim_{M\to\infty} \frac{1}{M}\hat{d}^{st} = \lim_{M\to\infty} \frac{1}{M}d^{st}. \tag{2.36}$$

Sparse Strata

We have

$$\begin{aligned}
\mathbb{E}\hat{d}^{st}_{ab} &= \frac{1}{N_k^2}\{\mathbb{E}X^{s}_{x|a}X^{t}_{y|a}X^{\bar{s}\bar{t}}_{xy|b} + \mathbb{E}X^{st}_{xy|a}X^{\bar{s}}_{x|b}X^{\bar{t}}_{y|b} - \mathbb{E}X^{st}_{xy|a}X^{\bar{s}\bar{t}}_{xy|b}\} \\
&= \frac{1}{N_k^2}\{\mathbb{E}X^{s}_{x|a}X^{t}_{y|a}\mathbb{E}X^{\bar{s}\bar{t}}_{xy|b} + \mathbb{E}X^{st}_{xy|a}\mathbb{E}X^{\bar{s}}_{x|b}X^{\bar{t}}_{y|b} - \mathbb{E}X^{st}_{xy|a}\mathbb{E}X^{\bar{s}\bar{t}}_{xy|b}\} \\
&= \frac{1}{N_k^2}\{(n_a n'_a \pi^{s}_{x|a}\pi^{t}_{y|a} + n_a\pi^{st}_{xy|a})\pi^{\bar{s}\bar{t}}_{xy|b} + \pi^{st}_{xy|a}(n_b n'_b \pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|b} + n_b\pi^{\bar{s}\bar{t}}_{xy|b}) - \pi^{st}_{xy|a}\pi^{\bar{s}\bar{t}}_{xy|b}\} \\
&= \frac{n_a n_b}{N_k^2}\{n'_a\pi^{s}_{x|a}\pi^{t}_{y|a}\pi^{\bar{s}\bar{t}}_{xy|b} + n'_b\pi^{st}_{xy|a}\pi^{\bar{s}\bar{t}}_{xy|b} + \pi^{st}_{xy|a}\pi^{\bar{s}\bar{t}}_{xy|b}\} \\
&= d^{st}_{ab}
\end{aligned}$$

and

$$\mathbb{E}\hat{d}^{st}_{abc} = \frac{1}{N_k^2}\mathbb{E}X^{st}_{xy|a}X^{\bar{s}}_{x|b}X^{\bar{t}}_{y|c} = \frac{1}{N_k^2}n_a n_b n_c \pi^{st}_{xy|a}\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|c} = d^{st}_{abc}.$$

By the Chebyshev law of large numbers we conclude

$$\frac{\sum_k \hat{d}^{st}}{K} \xrightarrow[p]{K\to\infty} \frac{\sum_k \mathbb{E}\hat{d}^{st}}{K} = \frac{\sum_k d^{st}}{K}$$

which was to be shown.

Large Strata:

As before, we consider the case $N \to \infty$ with $N\alpha_{ik} = n_{ik}$ and $1 > \alpha_{ik} > 0$. We compute

$$\begin{aligned}
d^{st}_{xy|ab}/N &= \frac{n_a n_b}{N_k^2 N}\{n'_a\pi^s_{x|a}\pi^t_{y|a}\pi^{\bar{s}\bar{t}}_{xy|b} + n'_b\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|b}\pi^{st}_{xy|a} + \pi^{\bar{s}\bar{t}}_{xy|a}\pi^{st}_{xy|b}\} \\
&= \frac{n_a n_b}{N^2}\frac{N_k^2}{N^2}\{\frac{n'_a}{N}\pi^s_{x|a}\pi^t_{y|a}\pi^{\bar{s}\bar{t}}_{xy|b} + \frac{n'_b}{n_b}\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|b}\pi^{st}_{xy|a} + \frac{1}{N}\pi^{\bar{s}\bar{t}}_{xy|a}\pi^{st}_{xy|b}\} \\
&\xrightarrow{N\to\infty} \frac{\alpha_a\alpha_b}{(\sum_i \alpha_i)^2}\{\alpha_a\pi^s_{x|a}\pi^t_{y|a}\pi^{\bar{s}\bar{t}}_{xy|b} + \alpha_b\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|b}\pi^{st}_{xy|a} + 0\cdot\pi^{\bar{s}\bar{t}}_{xy|a}\pi^{st}_{xy|b} \\
&= \frac{\alpha_a\alpha_b}{(\sum_i \alpha_i)^2}\{\alpha_a\pi^s_{x|a}\pi^t_{y|a}\pi^{\bar{s}\bar{t}}_{xy|b} + \alpha_b\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|b}\pi^{st}_{xy|a}\},
\end{aligned}$$

$$\begin{aligned}
\hat{d}^{st}_{xy|ab}/N &= \frac{1}{N_k^2 N}\{X^s_{x|a}X^t_{y|a}X^{\bar{s}\bar{t}}_{xy|b} + X^{st}_{xy|a}X^{\bar{s}}_{x|b}X^{\bar{t}}_{y|b} - X^{st}_{xy|a}X^{\bar{s}\bar{t}}_{xy|b}\} \\
&= \frac{n_a n_b}{N^2}\frac{N^2}{N_k^2}\left\{\frac{n_a}{N}\frac{X^s_{x|a}}{n_a}\frac{X^t_{y|a}}{n_a}\frac{X^{\bar{s}\bar{t}}_{xy|b}}{n_b} + \frac{n_b}{N}\frac{X^{st}_{xy|a}}{n_b}\frac{X^{\bar{s}}_{x|b}}{n_a}\frac{X^{\bar{t}}_{y|b}}{n_b} - \frac{1}{N}\frac{X^{st}_{xy|a}}{n_a}\frac{X^{\bar{s}\bar{t}}_{xy|b}}{n_b}\right\} \\
&\xrightarrow{N\to\infty} \frac{\alpha_a\alpha_b}{(\sum_i \alpha_i)^2}\{\alpha_a\pi^s_{x|a}\pi^t_{y|a}\pi^{\bar{s}\bar{t}}_{xy|b} + \alpha_b\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|b}\pi^{st}_{xy|a} + 0\cdot\pi^{\bar{s}\bar{t}}_{xy|a}\pi^{st}_{xy|b} \\
&= \frac{\alpha_a\alpha_b}{(\sum_i \alpha_i)^2}\{\alpha_a\pi^s_{x|a}\pi^t_{y|a}\pi^{\bar{s}\bar{t}}_{xy|b} + \alpha_b\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|b}\pi^{st}_{xy|a}\},
\end{aligned}$$

$$d^{st}_{abc}/N = \frac{n_a n_b n_c}{N_k^2 N}\pi^{st}_{xy|a}\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|c} = \frac{n_a n_b n_c}{N^3}\frac{N^2}{N_k^2}\pi^{st}_{xy|a}\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|c}$$
$$\stackrel{N\to\infty}{\longrightarrow} \frac{\alpha_a\alpha_b\alpha_c}{(\sum_i \alpha_i)^2}\pi^{st}_{xy|a}\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|c}$$

and

$$\hat{d}^{st}_{abc}/N = \frac{1}{N_k^2 N}X^{st}_{xy|a}X^{\bar{s}}_{x|b}X^{\bar{t}}_{y|c} = \frac{n_a n_b n_c}{N^3}\frac{N^2}{N_k^2}\frac{X^{st}_{xy|a}}{n_a}\frac{X^{\bar{s}}_{x|b}}{n_b}\frac{X^{\bar{t}}_{y|c}}{n_c}$$
$$\stackrel{N\to\infty}{\longrightarrow} \frac{\alpha_a\alpha_b\alpha_c}{(\sum_i \alpha_i)^2}\pi^{st}_{xy|a}\pi^{\bar{s}}_{x|b}\pi^{\bar{t}}_{y|c}.$$

We just showed that $\lim_{N\to\infty}\frac{1}{N}d^{st}_{ab} = \lim_{N\to\infty}\frac{1}{N}\hat{d}^{st}_{ab}$ and $\lim_{N\to\infty}\frac{1}{N}d^{st}_{abc} = \lim_{N\to\infty}\frac{1}{N}\hat{d}^{st}_{abc}$, which is (2.36), thus $U_{xy|abb}$ and $U_{xy|abc}$ are dually consistent.

Derivation of Covariance Estimators for the Generalised Log Odds Ratio Estimators

The common log odds ratio $\log\Psi^x_{ab}$ can be estimated by L^x_{ab} but more efficiently by the generalised estimator $\bar{L}^x_{ab}$ defined by (2.4) on page 48. The (co)variances for the generalised estimator are computed from the (co)variances of the estimators L^x_{ab}, because $\bar{L}^x_{ab}$ is a linear combination of the L^x_{ab} and so are the covariances. We prove now formula (2.12) on page 52, the (co)variance estimator for the generalised estimator $\bar{L}^x_{ab}$.

$$\mathrm{Cov}(\bar{L}^x_{ab},\bar{L}^y_{cd}) = \mathrm{Cov}(1/r\sum_{h=1}^{r}L^x_{ah} - L^x_{bh}, 1/r\sum_{i=1}^{r}L^y_{ci} - L^y_{di})$$
$$= 1/r^2\sum_{h,i}\{\mathrm{Cov}(L^x_{ah},L^y_{ci}) + \mathrm{Cov}(L^x_{bh},L^y_{di}) - \mathrm{Cov}(L^x_{ah},L^y_{di}) - \mathrm{Cov}(L^x_{bh},L^y_{ci})\}$$
$$= 1/r^2\sum_{i}\{\mathrm{Cov}(L^x_{ai},L^y_{ci}) + \mathrm{Cov}(L^x_{bi},L^y_{di}) - \mathrm{Cov}(L^x_{ai},L^y_{di}) - \mathrm{Cov}(L^x_{bi},L^y_{ci})\}$$

$$
\begin{aligned}
&+ 1/r^2 \sum_{h \neq i} \{\mathrm{Cov}(L^x_{ah}, L^y_{ci}) + \mathrm{Cov}(L^x_{bh}, L^y_{di}) - \mathrm{Cov}(L^x_{ah}, L^y_{di}) - \mathrm{Cov}(L^x_{bh}, L^y_{ci})\} \\
&= 1/r^2 \{U_{xy|+ac} + U_{xy|+bd} - U_{xy|+ad} - U_{xy|+bc}\} \\
&+ 1/r^2 \sum_{h \neq i} \{\mathrm{Cov}(L^x_{ah}, L^y_{ci}) + \mathrm{Cov}(L^x_{bh}, L^y_{di}) - \mathrm{Cov}(L^x_{ah}, L^y_{di}) - \mathrm{Cov}(L^x_{bh}, L^y_{ci})\}
\end{aligned}
$$

We express $\sum_{h \neq i} \mathrm{Cov}(L^x_{ah}, L^y_{ci})$ as

$$
\begin{aligned}
\sum_{\substack{h,i \\ h \neq i}} \mathrm{Cov}(L^x_{ah}, L^y_{ci}) &= \sum_{\substack{h \\ (i=a)}} \mathrm{Cov}(L^x_{ah}, L^y_{ca}) + \sum_{\substack{i \\ (h=c)}} \mathrm{Cov}(L^x_{ac}, L^y_{ci}) \\
&\quad - \mathrm{Cov}(L^x_{ac}, L^y_{ca}) + \sum_{\substack{h,i \\ c \neq h \neq i \neq a}} \mathrm{Cov}(L^x_{ah}, L^y_{ci}) \\
&= \sum_{\substack{h \\ (i=a)}} \mathrm{Cov}(L^x_{ah}, L^y_{ca}) + \sum_{\substack{i \\ (h=c)}} \mathrm{Cov}(L^x_{ac}, L^y_{ci}) - \mathrm{Cov}(L^x_{ac}, L^y_{ca}).
\end{aligned}
\tag{2.37}
$$

The second equality follows from $\mathrm{Cov}(L^x_{ai}, L^y_{ck}) = 0$ for distinct indices a, i, c, k, because the rows are independent. Thus

$$
\sum_{\substack{h,i \\ h \neq i}} \widehat{\mathrm{Cov}}(L^x_{ah}, L^y_{ci}) = -U_{xy|a+c} - U_{xy|ca+} + U_{xy|ac}. \tag{2.38}
$$

It follows

$$
\widehat{\mathrm{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{cd}) = \frac{1}{r^2} \{U^+_{xy|ac} - U^+_{xy|ad} - U^+_{xy|bc} + U^+_{xy|bd}\} \tag{2.39}
$$

with

$$
U^+_{xy|ac} = \begin{cases} U_{xy|a++} = \sum_{h,i} \mathrm{Cov}(L^x_{ah}, L^y_{ai}) & , a = c \\ U_{xy|+ac} - U_{xy|a+c} - U_{xy|ca+} + U_{xy|ac} & , a \neq c \end{cases}
$$

For non-distinct indices a, b, c, d we obtain the following sub-cases

$$\widehat{\text{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{ac}) = \frac{1}{r^2}\{U^+_{xy|a++} - U^+_{xy|ac} - U^+_{xy|ba} + U^+_{xy|bc}\} \tag{2.40}$$

and

$$\widehat{\text{Cov}}(\bar{L}^x_{ab}, \bar{L}^y_{ab}) = \frac{1}{r^2}\{U_{xy|a++} - U^+_{xy|ab} - U^+_{xy|ba} + U_{xy|b++}\}. \tag{2.41}$$

2.8.2 Proof of Influence Measure

We want to show that the influence measure defined by (2.17) on page 68 with

$$\hat{\boldsymbol{\beta}} = (\bar{\mathbf{L}}^T_{12}, \ldots, \bar{\mathbf{L}}^T_{1r})^T$$

and $\bar{\mathbf{L}}_{ab} = (\bar{L}^1_{ab}, \ldots, \bar{L}^J_{ab})^T$ is equal to the influence measure replacing $\hat{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\beta}}'$, where $\hat{\boldsymbol{\beta}}'$ contains any $r-1$ independent vectors $\bar{\mathbf{L}}_{ab}$ with $a \neq b; a, b \in \{1, \ldots, r\}$. Every $\bar{\mathbf{L}}_{ab}, b > a > 1$ can be re-expressed in terms of the vectors $\bar{\mathbf{L}}_{1b}$ in $\hat{\boldsymbol{\beta}}$ as

$$\bar{\mathbf{L}}_{ab} = \bar{\mathbf{L}}_{1b} - \bar{\mathbf{L}}_{1a}, b > a > 1.$$

Therefore $\hat{\boldsymbol{\beta}}' = \mathbf{C}\hat{\boldsymbol{\beta}}$, with an invertible matrix $\mathbf{C}$ having only elements -1, 0, and $+1$. If $\mathbf{C}$ is not invertible, then $\hat{\boldsymbol{\beta}}'$ does not contain $r-1$ independent vectors. This relationship holds for both $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{[d]}$.

Define $\mathbf{y} := \hat{\boldsymbol{\beta}} - \mathbb{E}\hat{\boldsymbol{\beta}}$ and it follows the covariance matrix can be written as $\text{Cov}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\mathbf{y}\mathbf{y}^T$, equivalently for $\hat{\boldsymbol{\beta}}'$. Also define $\mathbf{x} := \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[d]}$ and $\mathbf{x}' := \hat{\boldsymbol{\beta}}' - \hat{\boldsymbol{\beta}}'_{[d]}$ and it follows $\mathbf{x}' = \mathbf{C}\mathbf{x}$ and $\mathbf{y}' = \mathbf{C}\mathbf{y}$. We have

$$\begin{aligned} p \cdot CD(\boldsymbol{\beta}')_{[d]} &= \mathbf{x}'^T(\mathbb{E}(\mathbf{y}'\mathbf{y}'^T))^{-1}\mathbf{x}' \\ &= (\mathbf{C}\mathbf{x})^T(\mathbb{E}((\mathbf{C}\mathbf{y})(\mathbf{C}\mathbf{y})^T))^{-1}\mathbf{C}\mathbf{x} \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{x}^T\mathbf{C}^T(\mathbf{C}\mathbb{E}(\mathbf{y}\mathbf{y}^T)\mathbf{C}^T)^{-1}\mathbf{C}\mathbf{x} \\
&= \mathbf{x}^T\mathbf{C}^T(\mathbf{C}^T)^{-1}\mathbb{E}(\mathbf{y}\mathbf{y}^T)^{-1}\mathbf{C}^{-1}\mathbf{C}\mathbf{x} \\
&= p \cdot CD(\boldsymbol{\beta}_{[d]})
\end{aligned}
$$

using the real covariance matrix. Replacing the real covariance matrix by the bootstrap estimate of covariance yields the same result.

Chapter 3

MH Estimators for Stratified Multiple Response Data with Two Independent Rows per Stratum

3.1 Introduction

For an ordinary 2×2 table the odds ratio is defined in terms of the four table probabilities formed by the two rows and two columns. For $r \times J$ tables, there are $\binom{r}{2}$ pairs of rows and $\binom{J}{2}$ pairs of columns defining $\binom{r}{2} \cdot \binom{J}{2}$ odds ratios. However, each of these odds ratios can be computed from the $(r-1) \times (J-1)$ *local odds ratio* defined as (Agresti 2002, p.55)

$$\Psi_{ij} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}},$$

where π_{ij} is the probability of selecting row i and column j. The local odds ratios form a non-unique minimal set of odds ratios.

Now we want to consider such (local) odds ratios for stratified multiple re-

sponse data. First we consider the case of K $2 \times J$ tables and then generalise to K $r \times J$ tables. We define the (local) odds ratio as

$$\Psi_{xy|abk} = \frac{\pi_{x|ak}\pi_{y|bk}}{\pi_{y|ak}\pi_{x|bk}}, \tag{3.1}$$

where $\pi_{x|ak}$ is the probability of a positive response of item $x = 1\ldots, J$, row $a = 1, \ldots, r$ and stratum $k = 1, \ldots, K$. As before, we assume a common odds ratio

$$\Psi_{xy|ab} = \Psi_{xy|ab1} = \Psi_{xy|ab2} = \cdots = \Psi_{xy|abK} \tag{3.2}$$

for all strata. The MH estimator (Mantel and Haenszel 1959) now has the following form

$$\hat{\Psi}_{xy|ab} = C_{xy|ab}/C_{yx|ab}, \tag{3.3}$$

where $C_{xy|ab} = \sum_{k=1}^{K} c_{xy|abk}$, $c_{xy|k} = X_{x|ak}X_{y|bk}/N_k$ and $N_k = \sum_{i=1}^{r} n_{ik}$. Let us also define $L_{xy|ab} = \log \hat{\Psi}_{xy|ab}$. For two rows we simply suppress indices $a = 1$ and $b = 2$, for example, we write $\Psi_{xy|k}$ instead of $\Psi_{xy|12k}$. The kth odds ratio $\Psi_{xy|abk}$ describes the conditional relationship between two items and two rows, whereas the kth odds ratio Ψ^{x}_{abk} defined by (2.2) on page 48 describes the conditional relationship between two rows and one item only.

In the next section (Sec. 3.2), we show that the MH estimator (3.3) is still dually consistent under the assumptions of a common odds ratio and independent rows. However, in general, the dually consistent covariance and variance estimators proposed by Greenland (1989) are not applicable anymore. Then in Section 3.3, we derive dually consistent variance and covariance estimators for $\hat{\Psi}_{xy}$ and L_{xy}. Section 3.4 derives generalised MH estimators for K $2 \times J$ tables, and Section 3.5 considers the generalised MH estimators for the extended case of K $r \times J$ tables. We focus in Section 3.6 on the UTI example to illustrate the newly defined

MH estimators and its dually consistent (co-)variance estimators. The last section (Sec. 3.7) finishes with a simulation study investigating the performance of the various estimators.

3.2 Dual Consistency of the Ordinary MH Estimator

Let us use the same notations from the previous chapter. From (2.20) on page 74 we have

$$EX_xX_y = nn'\pi_x\pi_y + n\pi^{11}_{xy}, \tag{3.4}$$

and we also recall the first two moments of the multinomial distribution

$$\begin{aligned} \mathbb{E}X &= n\pi \\ \mathbb{E}X^2 &= nn'\pi^2 + n\pi. \end{aligned} \tag{3.5}$$

As before, we consider two kinds of asymptotics, the "large-stratum" limiting model (or model I), where the row totals n_k grow without bound, and the "sparse-data" limiting model (or model II) with bounded stratum margins N_k, where K grows with the sample size ($K \to \infty$).

Theorem 3.2.1. *The common Mantel-Haenszel estimator $\hat{\Psi}_{xy}$ in (3.3) is also dually consistent for the sampling model comprising of independent rows of multiple responses under the common odds-ratio assumption.*

Proof. Sparse-Data: From

$$\pi_{x|1k}\pi_{y|2k} = \Psi_{xy}\pi_{y|1k}\pi_{x|2k}, \tag{3.6}$$

which follows from the the common odds ratio assumption (3.2), we derive

$$\begin{aligned}\mathbb{E}\omega_{xy|k} = \mathbb{E}(c_{xy|k} - \Psi_{xy}c_{yx|k}) =& \mathbb{E}c_{xy|k} - \Psi_{xy}\mathbb{E}c_{yx|k} \\ =& \{\mathbb{E}X_{x|1k}\mathbb{E}X_{y|2k} - \Psi_{xy}\mathbb{E}X_{y|1k}\mathbb{E}X_{x|2k}\}/N_k \\ =& \{n_{1k}n_{2k}\pi_{x|1k}\pi_{y|2k} - \Psi_{xy}n_{1k}n_{2k}\pi_{y|1k}\pi_{x|2k}\}/N_k \\ =& \{n_{1k}n_{2k}(\pi_{x|1k}\pi_{y|2k} - \pi_{x|1k}\pi_{y|2k})\}/N_k = 0 \end{aligned} \tag{3.7}$$

with $\omega_{xy|abk} := c_{xy|abk} - \Psi_{xy|ab}c_{yx|abk}$ and $\Omega_{xy|ab} := \sum_k \omega_{xy|abk}$.
We can write

$$\begin{aligned}\hat{\Psi}_{xy} - \Psi_{xy} &= \frac{\sum_{k=1}^K c_{xy|k} - \Psi_{xy}c_{yx|k}}{\sum_{k=1}^K c_{yx|k}} \\ &= \frac{\sum_{k=1}^K (c_{xy|k} - \Psi_{xy}c_{yx|k})/K}{\sum_{k=1}^K c_{yx|k}/K} \\ &= \frac{\sum_{k=1}^K \omega_{xy|k}/K}{\sum_{k=1}^K c_{yx|k}/K} = \frac{\Omega_{xy}/K}{C_{yx}/K}. \end{aligned} \tag{3.8}$$

The term $c_{xy|k}$ is a bounded random variable under model II, hence, the variance of C_{xy} is $o(K^2)$ and Theorem 2.8.2 states $(\Omega_{xy} - \mathbb{E}\Omega_{xy})/K \to_p 0$. By (3.7) the expression reduces to $\Omega_{xy}/K \to_p 0$, that is, the numerator of $\hat{\Psi}_{xy} - \Psi_{xy}$ in (3.8) converges to zero in probability. Applying the Chebyshev weak law of large numbers again to the denominator yields

$$\sum_{k=1}^K c_{xy|k}/K \overset{K\to\infty}{\longrightarrow}_p \lim_{K\to\infty} \sum_{k=1}^K \mathbb{E}(c_{xy|k})/K < \infty. \tag{3.9}$$

This limit is finite and nonzero. Thus, we conclude $\hat{\Psi}_{xy} - \Psi_{xy} \to_p 0$ by Slutsky's theorem.

Large-Stratum: Let us consider the case $N \to \infty$ with $N\alpha_{ak} = n_{ak}$ and $0 <$

$\alpha_{ak} < 1$, that is, as N approaches infinity the number of subjects n_{ak}, for all rows a and strata k, also approaches infinity. Note $N_k = n_{1k} + n_{2k} = N \sum_i \alpha_{ik}$.

Generally, for the term $\sum_{k=1}^{K} \mathbb{E}c_{xy|k}/N$ we derive

$$\begin{aligned}
\mathbb{E}C_{xy|k}/N &= \sum_{k=1}^{K} \mathbb{E}c_{xy|k}/N = \sum_{k=1}^{K} \mathbb{E}X_{x|1k}\mathbb{E}X_{y|2k}/(N_k N) \\
&= \sum_{k=1}^{K} \frac{n_{1k}n_{2k}}{N_k N}\pi_{x|1k}\pi_{y|2k} = \sum_{k=1}^{K} \frac{n_{1k}n_{2k}}{NN}\frac{N}{N_k}\pi_{x|1k}\pi_{y|2k} \\
&\overset{N\to\infty}{\longrightarrow} \sum_{k=1}^{K} \alpha_{1k}\alpha_{2k}(\sum_i \alpha_{ik})^{-1}\pi_{x|1k}\pi_{y|2k} = \sum_{k=1}^{K}(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{x|1k}\pi_{y|2k} < \infty, \qquad (3.10)
\end{aligned}$$

also

$$\begin{aligned}
C_{xy|k}/N &= \sum_{k=1}^{K} c_{xy|k}/N = \sum_{k=1}^{K} X_{x|1k}X_{y|2k}/(N_k N) \\
&= \sum_{k=1}^{K} \frac{n_{1k}n_{2k}}{N_k N}\frac{X_{x|1k}}{n_1}\frac{X_{y|2k}}{n_2} = \sum_{k=1}^{K} \frac{n_{1k}n_{2k}}{NN}\frac{N}{N_k}\frac{X_{x|1k}}{n_1}\frac{X_{y|2k}}{n_2} \\
&\overset{N\to\infty}{\longrightarrow}_p \sum_{k=1}^{K} \alpha_{1k}\alpha_{2k}(\sum_i \alpha_{ik})^{-1}\pi_{x|1k}\pi_{y|2k} = \sum_{k=1}^{K}(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{x|1k}\pi_{y|2k}. \qquad (3.11)
\end{aligned}$$

In the following notation M can stand for either N or K. We showed

$$\lim_M C_{xy}/M = \lim_M \mathbb{E}C_{xy}/M < \infty \text{ with } M \in \{K, N\}. \qquad (3.12)$$

We also have

$$\lim_M \mathbb{E}C_{xy}/M = \Psi_{xy} \lim_M \mathbb{E}C_{yx}/M \text{ with } M \in \{K, N\} \qquad (3.13)$$

from property (3.6). By Slutsky's theorem, (3.12) and (3.13)

$$\hat{\Psi}_{xy} = \frac{C_{xy}}{C_{yx}} = \frac{C_{xy}/N}{C_{yx}/N} \overset{N\to\infty}{\longrightarrow}_p \frac{\lim_N \mathbb{E}C_{xy}/N}{\lim_N \mathbb{E}C_{yx}/N} = \Psi_{xy}\frac{\lim_N \mathbb{E}C_{yx}/N}{\lim_N \mathbb{E}C_{yx}/N} = \Psi_{xy}.$$

We could have used the same argument for the sparse data case, replacing N by K. It follows that $\hat{\Psi}_{xy}$ is dually consistent.

□

3.3 Dually Consistent Covariance and Variance Estimators

In this section, we derive dually consistent estimators for $\text{Var}(\hat{\Psi}_{xy})$, $\text{Cov}(\hat{\Psi}_{xy}, \hat{\Psi}_{xz})$, $\text{Cov}(\hat{\Psi}_{xy}, \hat{\Psi}_{wz})$ and for $\text{Var}(L_{xy})$, $\text{Cov}(L_{xy}, L_{xz})$, $\text{Cov}(L_{xy}, L_{wz})$.

3.3.1 Asymptotic Covariances and Variances

Using a similar argument as in Subsection 2.8.1 on page 76, we can derive a formula for the asymptotic covariances under both limiting models. We obtain similarly to equation (2.30)

$$\lim_{M\to\infty} M\cdot\text{Cov}^a(\hat{\Psi}_{xy|ab}, \hat{\Psi}_{wz|cd}) = \frac{\lim_{N\to\infty} \frac{1}{M}\sum_{k=1}^K \text{Cov}(\omega_{xy|abk}, \omega_{wz|cdk})}{[\lim_{N\to\infty} \frac{1}{M}\sum_{k=1}^K \mathbb{E}c_{yx|abk}][\lim_{N\to\infty} \frac{1}{M}\sum_{k=1}^K \mathbb{E}c_{zw|cdk}]} \tag{3.14}$$

with $M \in \{N, K\}$. For (3.14) to also be valid under the "large stratum" limiting model, we must show that $(\sqrt{N}\cdot\Omega_{xy|ab}/N)^2$ is uniformly integrable. Note that $\text{Cov}(\Omega_{xy|ab}, \Omega_{wz|cd}) = \sum_{k=1}^K \text{Cov}(\omega_{xy|abk}, \Omega_{wz|cdk})$ under independence of strata.

First we compute the kth variance $\text{Var}(\omega_{xy|abk})$

$$
\begin{aligned}
&\frac{N_k^2}{n_{1k}n_{2k}}\text{Var}(\omega_{xy|abk}) = \frac{N_k^2}{n_{1k}n_{2k}}\text{Var}(c_{xy|k} - \Psi c_{yx|k}) \\
&= \frac{N_k^2}{n_{1k}n_{2k}}\left[\mathbb{E}(c_{xy|k} - \Psi c_{yx|k})^2 - (\mathbb{E}(c_{xy|k} - \Psi c_{yx|k}))^2\right] \\
&= \frac{N_k^2}{n_{1k}n_{2k}}\mathbb{E}(c_{xy} - \Psi_{xy}c_{yx})^2 = \frac{N_k^2}{n_{1k}n_{2k}}\left[\mathbb{E}c_{xy}^2 - 2\Psi c_{xy}c_{yx} + \Psi^2 c_{yx}^2\right] \\
&= \frac{1}{n_{1k}n_{2k}}\left[\mathbb{E}X_{x|1}^2\mathbb{E}X_{y|2}^2 - 2\Psi\mathbb{E}X_{x|1}X_{y|1}\mathbb{E}X_{x|2}X_{y|2} + \Psi^2\mathbb{E}X_{y|1}^2\mathbb{E}X_{x|2}^2\right] \\
&= (\pi_{x|1} + n_1'\pi_{x|1}^2)(\pi_{y|2} + n_2'\pi_{y|2}^2) \\
&- 2\Psi(n_1'\pi_{x|1}\pi_{y|1} + \pi_{xy|1})(n_2'\pi_{x|2}\pi_{y|2} + \pi_{xy|2}) \\
&+ \Psi^2(\pi_{y|1} + n_1'\pi_{y|1}^2)(\pi_{x|2} + n_2'\pi_{x|2}^2) \\
&= (\pi_{x|1}\pi_{y|2} + n_1'\pi_{x|1}^2\pi_{y|2} + n_2'\pi_{x|1}\pi_{y|2}^2 + n_1'n_2'\pi_{x|1}^2\pi_{y|2}^2) \\
&- 2\Psi(n_1'n_2'\pi_{x|1}\pi_{y|1}\pi_{x|2}\pi_{y|2} + n_1'\pi_{x|1}\pi_{y|1}\pi_{xy|2} + n_2'\pi_{x|2}\pi_{y|2}\pi_{xy|1} + \pi_{xy|1}\pi_{xy|2}) \\
&+ \Psi^2(\pi_{y|1}\pi_{x|2} + n_1'\pi_{y|1}^2\pi_{x|2} + n_2'\pi_{y|1}\pi_{x|2}^2 + n_1'n_2'\pi_{y|1}^2\pi_{x|2}^2) \\
&= (\pi_{x|1}\pi_{y|2} + n_1'\pi_{x|1}^2\pi_{y|2} + n_2'\pi_{x|1}\pi_{y|2}^2) \\
&- 2\Psi(n_1'\pi_{x|1}\pi_{y|1}\pi_{xy|2} + n_2'\pi_{x|2}\pi_{y|2}\pi_{xy|1} + \pi_{xy|1}\pi_{xy|2}) \\
&+ \Psi^2(\pi_{y|1}\pi_{x|2} + n_1'\pi_{y|1}^2\pi_{x|2} + n_2'\pi_{y|1}\pi_{x|2}^2) \\
&= \frac{N_k^2}{n_{1k}n_{2k}}\{v_{xy|k}^1 - 2\Psi v_{xy|k}^2 + \Psi^2 v_{xy|k}^3\} \qquad (3.15)
\end{aligned}
$$

The third equality follows from (3.7), the fifth by applying (3.5) and (3.4), and the second to last by property (3.6).

In the next step, we compute the "large stratum" limiting variance $V = \lim_{N\to\infty} N\cdot \text{Var}^a(\Omega_{xy|ab}/N)$ by applying the delta method (Theorem 2.8.4). In Appendix A on page 289, we show that

$$\lim_{N\to\infty} N \cdot \text{Var}^a(\hat{\Psi}_{xy|ab}) = \frac{\lim_{N\to\infty}\sum_{k=1}^K \frac{1}{N}\text{Var}^a(\omega_{xy|k})}{[\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^K \mathbb{E}c_{yx|k}]^2}$$

$$= \frac{\sum_k \frac{(\sum_i \alpha_{ik}^{-1})^{-2}}{\alpha_{1k}}[\pi_{x|1k}\pi_{y|2k}^2 + \Psi^2\pi_{y|1k}\pi_{x|2k}^2 - 2\Psi\pi_{C|a}\pi_{x|2k}\pi_{y|2k}]}{(\sum_k(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{y|1k}\pi_{x|2k})^2}$$
$$+ \frac{\sum_k \frac{(\sum_i \alpha_{ik}^{-1})^{-2}}{\alpha_{2k}}[\pi_{x|1k}^2\pi_{y|2k} + \Psi^2\pi_{y|1k}^2\pi_{x|2k} - 2\Psi\pi_{C|b}\pi_{x|1k}\pi_{y|1k}]}{(\sum_k(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{y|1k}\pi_{x|2k})^2}. \tag{3.16}$$

under the common odds ratio assumption.

By (3.14)

$$\lim_{N\to\infty} N \cdot \mathrm{Var}^a(\hat{\Psi}_{xy}) = \frac{\lim_{N\to\infty}\sum_{k=1}^K \frac{1}{N}\mathrm{Var}(\omega_{xy|k})}{[\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^K \mathbb{E}c_{yx|k}]^2}. \tag{3.17}$$

We compute $\lim_{N\to\infty}\sum_k \mathrm{Var}(\omega_{xy|k})/N$ using (3.15)

$$\begin{aligned}
&\sum_k \mathrm{Var}(c_{xy|k} - \Psi c_{yx|k})/N \\
&= \sum_k \frac{n_1 n_2}{N_k^2 N}\{v_{xy|k}^1 - 2\Psi v_{xy|k}^2 + \Psi^2 v_{xy|k}^3\} \\
&= \sum_k \frac{n_1 n_2}{N_k^2 N}(\pi_{x|1}\pi_{y|2} + n_1'\pi_{x|1}^2\pi_{y|2} + n_2'\pi_{x|1}\pi_{y|2}^2) \\
&- \sum_k \frac{n_1 n_2}{N_k^2 N}2\Psi(n_1'\pi_{x|1}\pi_{y|1}\pi_{xy|2} + n_2'\pi_{x|2}\pi_{y|2}\pi_{xy|1} + \pi_{xy|1}\pi_{xy|2}) \\
&+ \sum_k \frac{n_1 n_2}{N_k^2 N}\Psi^2(\pi_{y|1}\pi_{x|2} + n_1'\pi_{y|1}^2\pi_{x|2} + n_2'\pi_{y|1}\pi_{x|2}^2) \\
&= \sum_k \frac{1}{N}\frac{n_1 n_2}{N^2}\frac{N^2}{N_k^2}\{\pi_{x|1}\pi_{y|2} - 2\Psi\pi_{xy|1}\pi_{xy|2} + \Psi^2\pi_{y|1}\pi_{x|2}\} \\
&+ \sum_k \frac{n_1 n_2 n_1'}{N^3}\frac{N^2}{N_k^2}\{\pi_{x|1}^2\pi_{y|2} + \Psi^2\pi_{y|1}^2\pi_{x|2} - 2\Psi\pi_{x|1}\pi_{y|1}\pi_{xy|2}\} \\
&+ \sum_k \frac{n_1 n_2 n_2'}{N^3}\frac{N^2}{N_k^2}\{\pi_{x|1}\pi_{y|2}^2 + \Psi^2\pi_{y|1}\pi_{x|2}^2 - 2\Psi\pi_{x|1}\pi_{y|1}\pi_{xy|2}\} \\
&\overset{N\to\infty}{\longrightarrow} 0 + \sum_k \frac{\alpha_1^2\alpha_2}{(\sum_i \alpha_{ik})^2}\{\pi_{x|1}^2\pi_{y|2} + \Psi^2\pi_{y|1}^2\pi_{x|2} - 2\Psi\pi_{x|1}\pi_{y|1}\pi_{xy|2}\} \\
&+ \sum_k \frac{\alpha_1\alpha_2^2}{(\sum_i \alpha_{ik})^2}\{\pi_{x|1}\pi_{y|2}^2 + \Psi^2\pi_{y|1}\pi_{x|2}^2 - 2\Psi\pi_{xy|1}\pi_{x|2}\pi_{y|2}\}.
\end{aligned} \tag{3.18}$$

If we substitute $\lim_{N\to\infty} \text{Var}(\omega_{xy|k})/N$ into (3.17), then equation (3.16) is identical to (3.17). Hence $\lim_{N\to\infty} \frac{1}{N}\text{Var}^a(\omega_{xy|k}\ /N)$ and $\lim_{N\to\infty} \frac{1}{N}\text{Var}(\omega_{xy|k})$ are also identical and so are $\lim_{N\to\infty} N \cdot \text{Var}^a(\Omega_{xy}\ /N)$and $\lim_{N\to\infty} N \cdot \text{Var}(\Omega_{xy}/N) = \lim_{N\to\infty} N \cdot \mathbb{E}(\Omega_{xy}/N)^2$. Then by Theorem 2.8.7, $(\sqrt{N} \cdot \Omega_{xy}/N)^2$ is uniformly integrable.

By Lemma 1 on page 79 and the uniform integrability of $(\sqrt{N} \cdot \Omega_{xy}/N)^2$, we conclude that $N \cdot \text{Cov}(\Omega_{xy|abk}\ /N, \Omega_{wz|cdk}/N)$ converges to $\lim_{N\to\infty} N \cdot \text{Cov}^a(\Omega_{xy|abk}\ /N, \Omega_{wz|cdk}/N)$, consequently formula (3.14) is indeed true for arbitrary indices x, y, a and b.

Remark 3.3.1. Appendix A on page 289 shows how costly the derivation of the "large-stratum" variance is when the the delta method is applied. In contrast, the computation of $\lim_{N\to\infty} \frac{1}{N}\ \text{Cov}(\Omega_{xy|abk}, \Omega_{wz|cdk})$ is quite cheap, provided Cov($\omega_{xy|abk}, \omega_{wz|cdk}$) is known. In the remainder of this chapter, we derive the limiting covariances only by (3.14), omitting the delta method.

Remark 3.3.2. We will use a trick, i.e. relabelling pairs of entries into a single sequence, to show that the odds ratio Ψ^x_{abk} defined in Chapter 2 is a special case of the odds ratio $\Psi_{xy|abk}$ defined in this chapter, this also applies to the underlying sampling models. Then we can conclude that the uniform integrability also holds in Chapter 2 (see pages 78-80), because we established it in this chapter.
Let us define $\tilde{\pi}_{2x-1|ak} := \pi_{x|ak}$, $\tilde{\pi}_{2x|ak} := 1 - \pi_{x|ak}$ and let $\text{Cov}(\tilde{X}_{2x-1|ak}, \tilde{X}_{2x|ak}) := -\tilde{\pi}_{2x-1|ak}\tilde{\pi}_{2x|ak}$ for $x = 1, \ldots, J$, $a = 1, \ldots, r$ and $k = 1, \ldots, K$. This ensures that $\tilde{X}_{2x-1|ak}$ and $\tilde{X}_{2x|ak}$ are simply the positive and negative responses of a binomial distribution, with probability of success $\pi_{x|ak}$ and probability of failure $1 - \pi_{x|ak}$. See Appendix B on page 296 for details. Also let the pairwise probabilities be defined as $\tilde{\pi}^{st}_{2x-1,2y-1|ak} := \pi^{st}_{xy|ak}$ with $s, t \in \{0, 1\}$. In a similar way, we set the higher order probabilities up to the Jth order, for example the fourth order probabilities $\tilde{\pi}^{stuv}_{2x-1,2y-1,2w-1,2z-1|ak} := \pi^{stuv}_{xywz|ak}$. It follows that r independent rows of pick any/c

variables, the sampling scheme of Chapter 2, is a special case of the sampling scheme of this chapter, assuming r independent rows of pick any$/2J$ variables. The odds ratios $\Psi_{xy|abk}$ based on $\{\tilde{\pi}_{x|ak}\}$ are identical to the odds ratios Ψ^x_{abk} based on $\{\pi_{x|ak}\}$. Hence Ψ^x_{abk} is a special case of $\Psi_{xy|abk}$.

We conclude that the uniform integrability of $(\sqrt{N} \cdot \Omega_{xy|ab}/N)^2$ also applies to $(\sqrt{N} \cdot \Omega_{x|ab}/N)^2$ (because $\Omega_{x|ab}$ is a special case of $\Omega_{xy|ab}$), a missing piece of the proof of equation (2.29) on page 78. In fact, all derivations of this chapter can be thought of as generalisations of those of Chapter 2.

3.3.2 A Dually Consistent Variance Estimator

We propose the following variance estimator of $\hat{\Psi}$

$$\widehat{\text{Var}}(\hat{\Psi}_{xy}) = \widehat{\Psi}^2_{xy} \widehat{\text{Var}}(L_{xy}) \tag{3.19}$$

with

$$\begin{aligned}
U_{xyy} := U_{xyxy} &:= \widehat{\text{Var}}(L_{xy}) = U^{old}_{xyy} + U^{add}_{xyy}, \\
U^{old}_{xyy} &:= \frac{\sum_k c_{xy} h_{xy}}{2C^2_{xy}} + \frac{\sum_k c_{yx} h_{yx}}{2C^2_{yx}} + \frac{\sum_k c_{xy} h_{yx} + c_{yx} h_{xy}}{2C_{xy} C_{yx}}, \\
U^{add}_{xyy} &:= -4 \frac{\sum_k X_{x|1} X_{y|1} X_{xy|2}/N^2_k + \sum_k X_{xy|1} X_{x|2} X_{y|2}/N^2_k}{2C_{xy} C_{yx}} \\
&\quad - \frac{\sum_k X_{xy|1}(X_{x|2} + X_{y|2})/N^2_k + \sum_k X_{xy|2}(X_{x|1} + X_{y|1})/N^2_k}{2C_{xy} C_{yx}} \\
&\quad + 4 \frac{\sum_k X_{xy|2} X_{xy|1}/N^2_k}{2C_{xy} C_{yx}}
\end{aligned} \tag{3.20}$$

and $h_{xy} := (X_{x|1} + X_{y|2})/N_k$. Equation (3.19) follows directly from the delta method. U^{old}_{xyy} is identical to the variance estimator suggested by Greenland (1989) for two rows of independent multinomials. $\widehat{\Psi}^2_{xy} \cdot U^{old}_{xyy}$ is also identical to the vari-

ance estimator $\widehat{\text{Var}}(\hat{\Psi}_{xy})$ suggested by Robins et al. (1986). Under the binomial and multinomial models, as considered by Greenland (1989), the variance estimator U^{old}_{xyy} is symmetric, that is, invariant under interchange of rows, of columns, and of rows and columns. However, due to U^{add}_{xyy}, the proposed variance estimator is only invariant under either interchange of rows or interchange of columns, but not invariant under interchange of rows and columns simultaneously. In fact, interchange of rows and columns simultaneously is very difficult to define under the multiple response sampling model.

Theorem 3.3.3. *U_{xyy} is a dually consistent estimator of Var$^a(L_{xy})$ and $\widehat{Var}(\hat{\Psi})$ is a dually consistent estimator of Var$^a(\hat{\Psi})$.*

Proof. By the Delta method

$$\text{Var}^a(L_{xy}) = \frac{1}{(\mathbb{E}^a\hat{\Psi})^2}\text{Var}^a(\hat{\Psi}) = \frac{1}{\Psi^2}\text{Var}^a(\hat{\Psi}_{xy}). \tag{3.21}$$

It is obvious from (3.21) that it is sufficient to show either of the two statements of the theorem, we show the dually consistency of $\widehat{\text{Var}}(\hat{\Psi}_{xy})$ $[= \Psi^2_{xy}U_{xyy}]$.

Sparse Data:

Note that all numerator terms $[\dots]_k$ of U_{xyy} are bounded random variables, hence, $\sum_k[\dots]_k/K$ converges as $K \to \infty$ to $\lim_{K\to\infty}\sum_k \mathbb{E}[\dots]_k/K$. Using (3.4),(3.5), (3.13) and Slutsky's theorem we have

$$\lim_{K\to\infty} K\cdot\widehat{\text{Var}}(\hat{\Psi}_{xy}) = \Psi^2_{xy}\cdot\lim_{K\to\infty} K\cdot U_{xyy} = \Psi^2_{xy}\cdot\lim_{K\to\infty} K\cdot(U^{old}_{xyy}+U^{add}_{xyy})$$
$$= \Psi^2\frac{\lim_K\sum_k \frac{1}{N_k^2K}\mathbb{E}X_{x|1}X_{y|2}(X_{x|1}+X_{y|2})}{2\lim_K(\mathbb{E}C_{xy}/K)^2} + \Psi^2\frac{\lim_K\sum_k \frac{1}{N_k^2K}\mathbb{E}X_{y|1}X_{x|2}(X_{y|1}+X_{x|2})}{2\lim_K(\mathbb{E}C_{yx}/K)^2}$$

$$
\begin{aligned}
&+ \Psi^2 \frac{\lim_K \sum_k \frac{1}{N_k^2 K}\{\mathbb{E}X_{x|1}X_{y|2}(X_{y|1}+X_{x|2}) + \mathbb{E}X_{y|1}X_{x|2}(X_{x|1}+X_{y|2})\}}{\lim_K (\mathbb{E}C_{xy}/K)(\mathbb{E}C_{yx}/K)} \\
&- 2\Psi^2 \frac{\lim_K \frac{1}{N_k^2 K}\{\sum_k \mathbb{E}X_{x|1}X_{y|1}X_{xy|2} + \mathbb{E}X_{xy|1}X_{x|2}X_{y|2} - \mathbb{E}X_{xy|2}X_{xy|1}\}}{\lim_K (\mathbb{E}C_{xy}/K)(\mathbb{E}C_{yx}/K)} \\
&- \Psi^2 \frac{\lim_K \sum_k \frac{1}{N_k^2 K}\{\mathbb{E}X_{xy|1}(X_{x|2}+X_{y|2}) + \mathbb{E}X_{xy|2}(X_{x|1}+X_{y|1})\}}{2\lim_K (\mathbb{E}C_{xy}/K)(\mathbb{E}C_{yx}/K)} \\
&= \frac{\lim_K \sum_k \frac{n_1 n_2}{N_k^2 K}\{\pi_{x|1}\pi_{y|2} + n_1'\pi_{x|1}^2\pi_{y|2} + \pi_{x|1}\pi_{y|2} + n_2'\pi_{x|1}\pi_{y|2}^2\}}{2(\mathbb{E}C_{yx}/K)^2} \\
&+ \frac{\Psi^2 \lim_K \sum_k \frac{n_1 n_2}{N_k^2}\{\pi_{y|1}\pi_{x|2} + n_1'\pi_{y|1}^2\pi_{x|2} + \pi_{y|1}\pi_{x|2} + n_2'\pi_{y|1}\pi_{x|2}^2\}}{2\lim_K (\mathbb{E}C_{yx}/K)^2} \\
&+ \frac{\lim_K \sum_k \frac{n_1 n_2}{N_k^2 K}\{\Psi^2 n_1'\pi_{y|1}^2\pi_{x|2} + \Psi\pi_{xy|1}\pi_{y|2} + \Psi^2 n_2'\pi_{y|1}\pi_{x|2}^2 + \Psi\pi_{xy|2}\pi_{x|1}\}}{2\lim_K (\mathbb{E}C_{yx}/K)^2} \\
&+ \frac{\lim_K \sum_k \frac{n_1 n_2}{N_k^2 K}\{n_1'\pi_{x|1}^2\pi_{y|2} + \Psi\pi_{xy|1}\pi_{x|2} + n_2'\pi_{x|1}\pi_{y|2}^2 + \Psi\pi_{xy|2}\pi_{y|1}\}}{2\lim_K (\mathbb{E}C_{yx}/K)^2} \\
&- \frac{\Psi \lim_K \sum_k \frac{n_1 n_2}{N_k^2 K}\{2n_1'\pi_{x|1}\pi_{y|1}\pi_{xy|2} + 2n_2'\pi_{x|2}\pi_{y|2}\pi_{xy|1} + 4\pi_{xy|1}\pi_{xy|2} - 2\pi_{xy|1}\pi_{xy|2}\}}{(\mathbb{E}C_{yx}/K)^2} \\
&- \frac{\lim_K \sum_k \frac{n_1 n_2}{N_k^2 K}\{\Psi\pi_{xy|1}\pi_{x|2} + \pi_{xy|1}\pi_{y|2} + \Psi\pi_{xy|2}\pi_{x|1} + \pi_{xy|2}\pi_{y|1}\}}{2\lim_K (\mathbb{E}C_{yx}/K)^2} \\
&= \frac{\lim_K \sum_k \frac{n_1 n_2}{N_k^2 K}\{\pi_{x|1}\pi_{y|2} + \Psi^2\pi_{y|1}\pi_{x|2} - 2\Psi\pi_{xy|1}\pi_{xy|2}\}}{\lim_K (\mathbb{E}C_{yx}/K)^2} \\
&+ \frac{\lim_K \sum_k \frac{n_1 n_2 n_1'}{N_k^2 K}\{\pi_{x|1}^2\pi_{y|2} + \Psi^2\pi_{y|1}^2\pi_{x|2} - 2\Psi\pi_{x|1}\pi_{y|1}\pi_{xy|2}\}}{\lim_K (\mathbb{E}C_{yx}/K)^2} \\
&+ \frac{\lim_K \sum_k \frac{n_1 n_2 n_2'}{N_k^2 K}\{\pi_{x|1}\pi_{y|2}^2 + \Psi^2\pi_{y|1}\pi_{x|2}^2 - 2\Psi\pi_{xy|1}\pi_{x|2}\pi_{y|2}\}}{\lim_K (\mathbb{E}C_{yx}/K)^2} \\
&= \frac{\lim_K \sum_k \{v_{1k} - 2\Psi v_{2k} + \Psi^2 v_{3k}\}}{\lim_K \mathbb{E}(C_{yx}/K)^2},
\end{aligned}
$$

which is identical to (3.17) with (3.18).

Large Stratum

$$
\begin{aligned}
&\lim_{N\to\infty} N \cdot \widehat{\mathsf{Var}}(\hat{\Psi}_{xy}) = \Psi_{xy}^2 \cdot \lim_{N\to\infty} N \cdot U_{xyy} = \Psi_{xy}^2 \cdot \lim_{N\to\infty} N \cdot (U_{xyy}^{old} + U_{xyy}^{add}) \\
&= \Psi^2 \frac{\lim_N \sum_k \frac{1}{N_k^2 N} X_{x|1}X_{y|2}(X_{x|1}+X_{y|2})}{2\lim_N \frac{1}{N^2} C_{xy}^2} + \Psi^2 \frac{\lim_N \sum_k \frac{1}{N_k^2 N} X_{y|1}X_{x|2}(X_{y|1}+X_{x|2})}{2\lim_N \frac{1}{N^2} C_{yx}^2}
\end{aligned}
$$

$$
\begin{aligned}
&+ \Psi^2 \frac{\lim_N \sum_k \frac{1}{N_k^2 N}\{X_{x|1}X_{y|2}(X_{y|1}+X_{x|2}) + X_{y|1}X_{x|2}(X_{x|1}+X_{y|2})\}}{2\lim_N \frac{1}{N^2}C_{xy}C_{yx}} \\
&- 2\Psi^2 \frac{\lim_N \frac{1}{N_k^2 N}\{\sum_k X_{x|1}X_{y|1}X_{xy|2} + X_{xy|1}X_{x|2}X_{y|2} - X_{xy|2}X_{xy|1}\}}{\lim_N \frac{1}{N^2}C_{xy}C_{yx}} \\
&- \Psi^2 \frac{\lim_N \frac{1}{N_k^2 N}\sum_k\{X_{xy|1}(X_{x|2}+X_{y|2}) + X_{xy|2}(X_{x|1}+X_{y|1})\}}{2\lim_N \frac{1}{N^2}C_{xy}C_{yx}} \\
&= \frac{\lim_N \sum_k \frac{n1n_2}{N_k^2}\left[\left\{\frac{n_1}{N}\frac{X_{x|1}^2}{n_1^2}\frac{X_{y|2}}{n_2} + \frac{n_2}{N}\frac{X_{x|1}}{n_1}\frac{X_{y|2}^2}{n_2^2}\right\} + \Psi^2\left\{\frac{n_1}{N}\frac{X_{y|1}^2}{n_1^2}\frac{X_{x|2}}{n_2} + \frac{n_2}{N}\frac{X_{y|1}}{n_1}\frac{X_{x|2}^2}{n_2^2}\right\}\right]}{2(\sum_{k=1}^K(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{y|1k}\pi_{x|2k})^2} \\
&+ \Psi \frac{\lim_N \sum_k \frac{n_1n_2}{N_k^2}\left\{\frac{n_1}{N}\frac{X_{x|1}}{n_1}\frac{X_{y|2}}{n_2}\frac{X_{y|1}}{n_1} + \frac{n_2}{N}\frac{X_{x|1}}{n_1}\frac{X_{y|2}}{n_2}\frac{X_{x|2}}{n_2} + \frac{n_1}{N}\frac{X_{y|1}}{n_1}\frac{X_{x|2}}{n_2}\frac{X_{x|1}}{n_1} + \frac{n_2}{N}\frac{X_{y|1}}{n_1}\frac{X_{x|2}}{n_2}\frac{X_{y|2}}{n_2}\right\}}{2(\sum_{k=1}^K(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{y|1k}\pi_{x|2k})^2} \\
&- 2\Psi \frac{\lim_N \sum_k \frac{n_1n_2}{N_k^2}\left\{\frac{n_1}{N}\frac{X_{x|1}}{n_1}\frac{X_{y|1}}{n_1}\frac{X_{xy|2}}{n_2} + \frac{n_2}{N}\frac{X_{xy|1}}{n_1}\frac{X_{x|2}}{n_2}\frac{X_{y|2}}{n_2} - \frac{1}{N}\frac{X_{xy|2}}{n_2}\frac{X_{xy|1}}{n_1}\right\}}{(\sum_{k=1}^K(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{y|1k}\pi_{x|2k})^2} \\
&- \Psi \frac{\lim_N \sum_k \frac{1}{N}\frac{n_1n_2}{N_k^2}\left\{\frac{X_{xy|1}}{n_1}\frac{X_{x|2}}{n_2} + \frac{X_{xy|1}}{n_1}\frac{X_{y|2}}{n_2} + \frac{X_{x|1}}{n_1}\frac{X_{xy|2}}{n_2} + \frac{X_{y|1}}{n_1}\frac{X_{xy|2}}{n_2}\right\}}{2(\sum_{k=1}^K(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{y|1k}\pi_{x|2k})^2} \\
&= \frac{\sum_k \frac{\alpha_1^2\alpha_2}{(\sum_i \alpha_{ik})^2}\{\pi_{x|1}\pi_{y|2}^2 + \sum_k \Psi^2\pi_{y|1}\pi_{x|2}^2\} + \frac{\alpha_1\alpha_2^2}{(\sum_i \alpha_{ik})^2}\{\pi_{x|1}^2\pi_{y|2} + \Psi^2\pi_{y|1}^2\pi_{x|2}\}}{(\sum_{k=1}^K(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{y|1k}\pi_{x|2k})^2} \\
&- 2\frac{\sum_k \frac{\alpha_1^2\alpha_2}{(\sum_i \alpha_{ik})^2}\Psi\pi_{xy|1}\pi_{x|2}\pi_{y|2} + \sum_k \frac{\alpha_1\alpha_2^2}{(\sum_i \alpha_{ik})^2}\Psi\pi_{xy|2}\pi_{x|1}\pi_{y|1} - 0}{(\sum_{k=1}^K(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{y|1k}\pi_{x|2k})^2} - 0
\end{aligned}
$$

which is identical to (3.16) or (3.17) with (3.18). □

3.3.3 Dually Consistent Covariance Estimators

We compute $\text{Cov}(\omega_{xy|k}, \omega_{xz|k})$

$$
\begin{aligned}
&\text{Cov}(\omega_{xy|k}, \omega_{xz|k}) = \text{Cov}(c_{xy} - \Psi_{xy}c_{yx}, c_{xz} - \Psi_{xz}c_{zx}) \\
&= \mathbb{E}c_{xy}c_{xz} - \Psi_{xy}c_{yx}c_{xz} - \Psi_{xz}c_{xy}c_{zx} + \Psi_{xy}\Psi_{xz}c_{yx}c_{zx} \\
&= \frac{1}{N^2}\{\mathbb{E}X_{x|1}^2\mathbb{E}X_{y|2}X_{z|2} - \Psi_{xy}\mathbb{E}X_{x|1}X_{y|1}\mathbb{E}X_{x|2}X_{z|2} \\
&- \Psi_{xz}\mathbb{E}X_{x|1}X_{z|1}\mathbb{E}X_{x|2}X_{y|2} + \Psi_{xy}\Psi_{xz}\mathbb{E}X_{y|1}X_{z|1}\mathbb{E}X_{x|2}^2\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n_1 n_2}{N^2}\{(\pi_{x|1} + n_1'\pi_{x|1}^2)(n_2'\pi_{y|2}\pi_{z|2} + \pi_{yz|2}) - \Psi_{xy}(n_1'\pi_{x|1}\pi_{y|1} + \pi_{xy|1})(n_2'\pi_{x|2}\pi_{z|2} + \pi_{xz|2}) \\
&- \Psi_{xz}(n_1'\pi_{x|1}\pi_{z|1} + \pi_{xz|1})(n_2'\pi_{x|2}\pi_{y|2} + \pi_{xy|2}) + \Psi_{xy}\Psi_{xz}(n_1'\pi_{y|1}\pi_{z|1} + \pi_{yz|1})(\pi_{x|2} + n_2'\pi_{x|2}^2)\} \\
&= \frac{n_1 n_2}{N^2}\{\pi_{x|1}\pi_{yz|2} - \Psi_{xy}\pi_{xy|1}\pi_{xz|2} - \Psi_{xz}\pi_{xz|1}\pi_{xy|2} + \Psi_{xy}\Psi_{xz}\pi_{yz|1}\pi_{x|2} \\
&+ n_1'(\pi_{x|1}^2\pi_{yz|2} - \Psi_{xy}\pi_{x|1}\pi_{y|1}\pi_{xz|2} - \Psi_{xz}\pi_{x|1}\pi_{z|1}\pi_{xy|2} + \Psi_{xy}\Psi_{xz}\pi_{y|1}\pi_{z|1}\pi_{x|2}) \\
&+ n_2'(\pi_{x|1}\pi_{y|2}\pi_{z|2} - \Psi_{xy}\pi_{xy|1}\pi_{x|2}\pi_{z|2} - \Psi_{xz}\pi_{xz|1}\pi_{x|2}\pi_{y|2} + \Psi_{xy}\Psi_{xz}\pi_{yz|1}\pi_{x|2}^2) \\
&+ n_1'n_2'(\pi_{x|1}^2\pi_{y|2}\pi_{z|2} - \Psi_{xy}\pi_{x|1}\pi_{y|1}\pi_{x|2}\pi_{z|2} - \Psi_{xz}\pi_{x|1}\pi_{z|1}\pi_{x|2}\pi_{y|2} + \Psi_{xy}\Psi_{xz}\pi_{y|1}\pi_{z|1}\pi_{x|2}^2)\} \\
&= \frac{n_1 n_2}{N^2}\{\pi_{x|1}\pi_{yz|2} - \Psi_{xy}\pi_{xy|1}\pi_{xz|2} - \Psi_{xz}\pi_{xz|1}\pi_{xy|2} + \Psi_{xy}\Psi_{xz}\pi_{yz|1}\pi_{x|2} \\
&+ n_1'(\pi_{x|1}^2\pi_{yz|2} - \Psi_{xy}\pi_{x|1}\pi_{y|1}\pi_{xz|2} - \Psi_{xz}\pi_{x|1}\pi_{z|1}\pi_{xy|2} + \Psi_{xy}\Psi_{xz}\pi_{y|1}\pi_{z|1}\pi_{x|2}) \\
&+ n_2'(\pi_{x|1}\pi_{y|2}\pi_{z|2} - \Psi_{xy}\pi_{xy|1}\pi_{x|2}\pi_{z|2} - \Psi_{xz}\pi_{xz|1}\pi_{x|2}\pi_{y|2} + \Psi_{xy}\Psi_{xz}\pi_{yz|1}\pi_{x|2}^2)\} \\
&= \frac{n_1 n_2}{N^2}\{\pi_{x|1}\pi_{yz|2} + n_1'\pi_{x|1}^2\pi_{yz|2} + n_2'\pi_{x|1}\pi_{y|2}\pi_{z|2}\} \\
&- \Psi_{xy}\frac{n_1 n_2}{N^2}\{\pi_{xy|1}\pi_{xz|2} + n_1'\pi_{x|1}\pi_{y|1}\pi_{xz|2} + n_2'\pi_{xy|1}\pi_{x|2}\pi_{z|2}\} \\
&- \Psi_{xz}\frac{n_1 n_2}{N^2}\{\pi_{xz|1}\pi_{xy|2} + n_1'\pi_{x|1}\pi_{z|1}\pi_{xy|2} + n_2'\pi_{xz|1}\pi_{x|2}\pi_{y|2}\} \\
&+ \Psi_{xy}\Psi_{xz}\frac{n_1 n_2}{N^2}\{\pi_{yz|1}\pi_{x|2} + n_1'\pi_{y|1}\pi_{z|1}\pi_{x|2} + n_2'\pi_{yz|1}\pi_{x|2}^2\} \\
&= \{v_{xyz|12,k} - \Psi_{xy}v_{xy,xz|k} - \Psi_{xz}v_{xz,xy|k} + \Psi_{xy}\Psi_{xz}v_{xyz|21,k}\} \qquad (3.22)
\end{aligned}
$$

with

$$
v_{xw,yz|k} = \frac{n_1 n_2}{N^2}\{\pi_{xw|1}\pi_{yz|2} + n_1'\pi_{x|1}\pi_{w|1}\pi_{yz|2} + n_2'\pi_{xw|1}\pi_{y|2}\pi_{z|2}\}, \qquad (3.23)
$$

$$
\begin{aligned}
v_{xyz|abk} &= v_{xyz|abk}^A + v_{xyz|abk}^B \ (a \neq b), \\
v_{xyz|abk}^A &= \frac{n_a n_b}{N^2}\pi_{yz|bk}\{\pi_{x|ak} + n_a'\pi_{x|ak}^2\} \ (a \neq b), \\
v_{xyz|abk}^B &= \frac{n_a n_b n_b'}{N^2}\pi_{x|ak}\pi_{y|bk}\pi_{z|bk} \ (a \neq b), \qquad (3.24)
\end{aligned}
$$

and V representing $\sum_k v_k$.

We rewrite (3.14) using (3.13) and (3.22) as

$$\begin{aligned}
&\lim_{M\to\infty} M \cdot \mathrm{Cov}^a(L_{xy}, L_{xz}) \\
&= \frac{\lim_M(V_{xyz|12,k}/M)}{\lim_M(\mathbb{E}C_{xy}/M)(\mathbb{E}C_{xz}/M)} - \frac{\lim_M(V_{xy,xz|k}/M)}{\lim_M(\mathbb{E}C_{yx}/M)(\mathbb{E}C_{xz}/M)} \\
&- \frac{\lim_M(V_{xz,xy|k}/M)}{\lim_M(\mathbb{E}C_{xy}/M)(\mathbb{E}C_{zx}/M)} + \frac{\lim_M(V_{xyz|21,k}/M)}{\lim_M(\mathbb{E}C_{yx}/M)(\mathbb{E}C_{zx}/M)}.
\end{aligned} \tag{3.25}$$

Similarly we compute

$$\begin{aligned}
&\mathrm{Cov}(\omega_{xy|k}, \omega_{wz|k}) = \mathrm{Cov}(c_{xy|k} - \Psi_{xy}c_{yx|k}, c_{wz|k} - \Psi_{wz}c_{zw|k}) \\
&= \mathbb{E}c_{xy}c_{wz} - \Psi_{xy}c_{yx}c_{wz} - \Psi_{wz}c_{xy}c_{zw} + \Psi_{xy}\Psi_{wz}c_{yx}c_{zw} \\
&= \frac{1}{N^2}\{\mathbb{E}X_{x|1}X_{w|1}\mathbb{E}X_{y|2}X_{z|2} - \Psi_{xy}\mathbb{E}X_{y|1}X_{w|1}\mathbb{E}X_{x|2}X_{z|2} \\
&- \Psi_{wz}\mathbb{E}X_{x|1}X_{z|1}\mathbb{E}X_{y|2}X_{w|2} + \Psi_{xy}\Psi_{wz}\mathbb{E}X_{y|1}X_{z|1}\mathbb{E}X_{x|2}X_{w|2}\} \\
&= \frac{n_1n_2}{N^2}\{(n_1'\pi_{x|1}\pi_{w|1} + \pi_{xw|1})(n_2'\pi_{y|2}\pi_{z|2} - \pi_{yz|2}) \\
&- \Psi_{xy}(n_1'\pi_{y|1}\pi_{w|1} + \pi_{yw|1})(n_2'\pi_{x|2}\pi_{z|2} + \pi_{xz|2}) \\
&- \Psi_{wz}(n_1'\pi_{x|1}\pi_{z|1} + \pi_{xz|1})(n_2'\pi_{y|2}\pi_{w|2} - \pi_{yw|2}) \\
&+ \Psi_{xy}\Psi_{wz}(n_1'\pi_{y|1}\pi_{z|1} + \pi_{yz|1})(n_2'\pi_{x|2}\pi_{w|2} + \pi_{xw|2})\} \\
&= \frac{n_1n_2}{N^2}\{\pi_{xw|1}\pi_{yz|2} - \Psi_{xy}\pi_{yw|1}\pi_{xz|2} - \Psi_{wz}\pi_{xz|1}\pi_{yw|2} + \Psi_{xy}\Psi_{wz}\pi_{yz|1}\pi_{xw|2} \\
&+ n_1'(\pi_{x|1}\pi_{w|1}\pi_{yz|2} - \Psi_{xy}\pi_{y|1}\pi_{w|1}\pi_{xz|2} - \Psi_{wz}\pi_{x|1}\pi_{z|1}\pi_{yw|2} + \Psi_{xy}\Psi_{wz}\pi_{y|1}\pi_{z|1}\pi_{xw|2}) \\
&+ n_2'(\pi_{xw|1}\pi_{y|2}\pi_{z|2} - \Psi_{xy}\pi_{yw|1}\pi_{x|2}\pi_{z|2} - \Psi_{wz}\pi_{xz|1}\pi_{y|2}\pi_{w|2} + \Psi_{xy}\Psi_{wz}\pi_{yz|1}\pi_{x|2}\pi_{w|2}) \\
&+ n_1'n_2'(\pi_{x|1}\pi_{w|1}\pi_{y|2}\pi_{z|2} - \Psi_{xy}\pi_{y|1}\pi_{w|1}\pi_{x|2}\pi_{z|2} \\
&- \Psi_{wz}\pi_{x|1}\pi_{z|1}\pi_{y|2}\pi_{w|2} + \Psi_{xy}\Psi_{wz}\pi_{y|1}\pi_{z|1}\pi_{x|2}\pi_{y|2})\} \\
&= \frac{n_1n_2}{N^2}\{\pi_{xw|1}\pi_{yz|2} - \Psi_{xy}\pi_{yw|1}\pi_{xz|2} - \Psi_{wz}\pi_{xz|1}\pi_{yw|2} + \Psi_{xy}\Psi_{wz}\pi_{yz|1}\pi_{xw|2} \\
&+ n_1'(\pi_{x|1}\pi_{w|1}\pi_{yz|2} - \Psi_{xy}\pi_{y|1}\pi_{w|1}\pi_{xz|2} - \Psi_{wz}\pi_{x|1}\pi_{z|1}\pi_{yw|2} + \Psi_{xy}\Psi_{wz}\pi_{y|1}\pi_{z|1}\pi_{xw|2})
\end{aligned}$$

$$+ n_2'(\pi_{xw|1}\pi_{y|2}\pi_{z|2} - \Psi_{xy}\pi_{yw|1}\pi_{x|2}\pi_{z|2} - \Psi_{wz}\pi_{xz|1}\pi_{y|2}\pi_{w|2} + \Psi_{xy}\Psi_{wz}\pi_{yz|1}\pi_{x|2}\pi_{w|2})\}$$
$$= \frac{n_1 n_2}{N^2}\{v_{xw,yz|k} - \Psi_{xy}v_{yw,xz|k} - \Psi_{wz}v_{xz,yw|k} + \Psi_{xy}\Psi_{wz}v_{yz,zw|k}\}. \tag{3.26}$$

Again, we rewrite (3.14) using (3.13) and (3.26) as

$$\begin{aligned}
&\lim_{M\to\infty} M \cdot \mathrm{Cov}^a(L_{xy}, L_{wz}) \\
&= \frac{\lim_M(V_{xw,yz}/M)}{\lim_M(\mathbb{E}C_{xy}/M)(\mathbb{E}C_{wz}/M)} - \frac{\lim_M(V_{yw,xz}/M)}{\lim_M(\mathbb{E}C_{yx}/M)(\mathbb{E}C_{wz}/M)} \\
&- \frac{\lim_M(V_{xz,yw}/M)}{\lim_M(\mathbb{E}C_{xy}/M)(\mathbb{E}C_{zw}/M)} + \frac{\lim_M(V_{yz,zw}/M)}{\lim_M(\mathbb{E}C_{yx}/M)(\mathbb{E}C_{zw}/M)}.
\end{aligned} \tag{3.27}$$

We propose the following estimators for $\mathrm{Cov}(L_{xy}, L_{xz})$, $\mathrm{Cov}(L_{xy}, L_{wz})$:

$$\begin{aligned}
U_{xyz} := U_{xyxz} &:= \widehat{\mathrm{Cov}}(L_{xy}, L_{xz}) \\
&= \frac{\hat{V}^A_{xyz|12k}}{C_{xy}C_{xz}} - \frac{\hat{V}_{xy,xz}}{C_{yx}C_{xz}} - \frac{\hat{V}_{xz,xy}}{C_{xy}C_{zx}} + \frac{\hat{V}^A_{xyz|21}}{C_{yx}C_{zx}} \\
&+ \frac{\hat{V}^B_{xyz|12k}}{3C_{xy}C_{xz}} + \frac{\hat{V}^B_{yxz|12} + \hat{V}^B_{zxy|21}}{3C_{yx}C_{xz}} + \frac{\hat{V}^B_{zxy|12} + \hat{V}^B_{yxz|21}}{3C_{xy}C_{zx}} + \frac{\hat{V}^B_{xyz|21k}}{3C_{yx}C_{zx}}
\end{aligned} \tag{3.28}$$

and

$$U_{xywz} := \widehat{\mathrm{Cov}}(L_{xy}, L_{wz}) = \frac{\hat{V}_{xw,yz}}{C_{xy}C_{wz}} - \frac{\hat{V}_{yw,xz}}{C_{yx}C_{wz}} - \frac{\hat{V}_{xz,yw}}{C_{xy}C_{zw}} + \frac{\hat{V}_{yz,xw}}{C_{yx}C_{zw}} \tag{3.29}$$

with

$$\hat{v}^A_{xyz|abk} = \frac{1}{N_k^2}X^2_{x|ak}X_{yz|bk} \tag{3.30}$$

$$\hat{v}^B_{xyz|abk} = \frac{1}{N_k^2}X_{x|ak}\{X_{y|bk}X_{z|bk} - X_{yz|bk}\} \tag{3.31}$$

$$\hat{v}_{xw,yz} = \frac{1}{N_k^2}\{X_{x|1}X_{w|1}X_{yz|2} + X_{xw|1}X_{y|2}X_{z|2} - X_{xw|1}X_{yz|2}\} \tag{3.32}$$

and $\hat{V}$ representing $\sum_k \hat{v}_k$.
The estimators $\widehat{\text{Cov}}(\hat{\Psi}_{xy}, \hat{\Psi}_{xz})$ and $\widehat{\text{Cov}}(\hat{\Psi}_{xy}, \hat{\Psi}_{wz})$ are computed as

$$\widehat{\text{Cov}}(\hat{\Psi}_{xy}, \hat{\Psi}_{xz}) = \Psi_{xy}\Psi_{xz}U_{xyz}$$
$$\widehat{\text{Cov}}(\hat{\Psi}_{xy}, \hat{\Psi}_{wz}) = \Psi_{xy}\Psi_{wz}U_{xywz}$$

by the delta method (Theorem 2.8.4).

We define U^{old}_{xyz} as an estimator consisting only of the second row of equation (3.28) and of equation (3.31), but which is amended to $\frac{1}{N_k^2}X_{x|ak}X_{y|bk}X_{z|bk}$. Then U^{old}_{xyz} is identical to the one proposed by Greenland (1989) for two independent rows of multinomials. Greenland (1989) did not define an estimator U_{xywz}, because $\text{Cov}(L_{xy}, L_{wz}) = 0$ (for distinct indices).

Theorem 3.3.4. *U_{xyz}, U_{xywz}, $\widehat{Cov}(\hat{\Psi}_{xy}, \hat{\Psi}_{xz})$, and $\widehat{Cov}(\hat{\Psi}_{xy}, \hat{\Psi}_{wz})$ are all dually consistent estimators.*

Proof. First we show that $\hat{v}^A_{xyz|abk}$ converges under both models to $v^A_{xyz|abk}$, then we also show $\hat{v}^B_{xyz|abk}$ converges to $\hat{v}^B_{xyz|abk}$ and $\hat{v}_{xw,yz}$ to $v_{xw,yz}$ under both limiting models.

Sparse Data

$$\begin{aligned}\lim_K(\hat{V}^A_{xyz|ab}/K) &= \lim_K(\mathbb{E}\hat{V}^A_{xyz|ab}/K) = \lim_K \frac{1}{K}\sum_k \frac{1}{N_k^2}\mathbb{E}X^2_{x|ak}\mathbb{E}X_{yz|bk} \\ &= \lim_K \frac{1}{K}\sum_k \frac{n_a n_b}{N_k^2}(n'_a\pi^2_{x|ak} + \pi_{x|a})\pi_{yz|bk} = \lim_K \frac{1}{K}\sum_k v^A_{xyz|ab} \\ &= \lim_K(V^A_{xyz|ab}/K)\end{aligned}$$

$$\lim_K(\hat{V}^B_{xyz|ab}/K) = \lim_K(\mathbb{E}\hat{V}^B_{xyz|ab}/K)$$

$$
\begin{aligned}
&= \lim_K \frac{1}{K} \sum_k \frac{1}{N_k^2} \mathbb{E} X_{x|ak} \{ \mathbb{E} X_{y|bk} X_{z|bk} - \mathbb{E} X_{yz|bk} \} \\
&= \lim_K \frac{1}{K} \sum_k \frac{n_a n_b}{N_k^2} \pi_{x|ak} \{ (n_b' \pi_{y|bk} \pi_{z|bk} + \pi_{yz|bk}) - \pi_{yz|bk} \} \\
&= \lim_K \frac{1}{K} \sum_k \frac{n_a n_b}{N_k^2} n_b' \pi_{x|ak} \pi_{y|bk} \pi_{z|bk} \\
&= \lim_K \frac{1}{K} \sum_k v_{xyz|ab}^B = \lim_K (V_{xyz|ab}^B / K)
\end{aligned}
$$

$$
\begin{aligned}
&\lim_K (\hat{V}_{xw,yz}/K) = \lim_K (\mathbb{E}\hat{V}_{xw,yz}/K) \\
&= \lim_K \frac{1}{K} \sum_k \frac{1}{N_k^2} \{ \mathbb{E} X_{x|1} X_{w|1} X_{yz|2} + \mathbb{E} X_{xw|1} X_{y|2} X_{z|2} - \mathbb{E} X_{xw|1} X_{yz|2} \} \\
&= \lim_K \frac{1}{K} \sum_k \frac{n_a n_b}{N_k^2} \{ (n_1' \pi_{x|1} \pi_{w|1} + \pi_{xw|1}) \pi_{yz|2} + \pi_{xw|1} (n_2' \pi_{y|2} \pi_{z|2} + \pi_{yz|2}) - \pi_{xw|1} \pi_{yz|2} \} \\
&= \lim_K \frac{1}{K} \sum_k \frac{n_a n_b}{N_k^2} \{ n_1' \pi_{x|1} \pi_{w|1} \pi_{yz|2} + n_2' \pi_{xw|1} \pi_{y|2} \pi_{z|2} + \pi_{xw|1} \pi_{yz|2} \} \\
&= \lim_K \frac{1}{K} \sum_k v_{xw,yz} = \lim_K (V_{xw,yz}/K)
\end{aligned}
$$

Large Stratum

$$
\begin{aligned}
\lim_N (\hat{V}_{xyz|ab}^A / N) &= \lim_N \frac{1}{N} \sum_k \frac{1}{N_k^2} X_{x|ak}^2 X_{yz|bk} \\
&= \lim_N \sum_k \frac{n_a n_b}{N_k^2} \left(\frac{n_a}{N} \left[\frac{X_{x|ak}}{n_a} \right]^2 + \frac{1}{N} \frac{X_{x|a}}{n_a} \right) \frac{X_{yz|bk}}{n_b} \\
&= \frac{1}{N} \sum_k \frac{\alpha_a \alpha_b}{(\sum_i \alpha_{ik})^2} (\alpha_a \pi_{x|ak}^2 + 0 \cdot \pi_{x|a}) \pi_{yz|bk}
\end{aligned}
$$

$$
\lim_N (V_{xyz|ab}^A / N) = \lim_N \sum_k \frac{n_a n_b}{N_k^2} \left\{ \frac{1}{N} \pi_{x|ak} + \frac{n_a'}{N} \pi_{x|ak}^2 \right\} \pi_{yz|bk}
$$

$$= \sum_k \frac{\alpha_a \alpha_b}{(\sum_i \alpha_{ik})^2}(0 \cdot \pi_{x|a} + \alpha_a \pi_{x|ak}^2)\pi_{yz|bk}$$

Hence, $\lim_N(\hat{V}^A_{xyz|ab}/N) = \lim_N(V^A_{xyz|ab}/N)$

$$\begin{aligned}
\lim_N(\hat{V}^B_{xyz|ab}/N) &= \lim_N \frac{1}{N}\sum_k \frac{1}{N_k^2}\{X_{x|ak}X_{y|bk}X_{z|bk} - X_{x|ak}X_{yz|bk}\} \\
&= \lim_N \sum_k \frac{n_a n_b}{N_k^2}\{\frac{n_b}{N}\frac{X_{x|ak}}{n_a}\frac{X_{y|bk}}{n_b}\frac{X_{z|bk}}{n_b} - \frac{1}{N}\frac{X_{x|ak}}{n_a}\frac{X_{yz|bk}}{n_b}\} \\
&= \sum_k \frac{\alpha_a \alpha_b}{(\sum_i \alpha_{ik})^2}\{\alpha_b \pi_{x|ak}\pi_{y|bk}\pi_{z|bk} - 0 \cdot \pi_{x|ak}\pi_{yz|bk}\}
\end{aligned}$$

$$\begin{aligned}
\lim_N(V^B_{xyz|ab}/N) &= \lim_N \frac{1}{N}\sum_k \frac{n_a n_b}{N_k^2} n_b' \pi_{x|ak}\pi_{y|bk}\pi_{z|bk} \\
&= \sum_k \frac{\alpha_a \alpha_b}{(\sum_i \alpha_{ik})^2}\alpha_b \pi_{x|ak}\pi_{y|bk}\pi_{z|bk}
\end{aligned}$$

It follows that $\lim_N(\hat{V}^B_{xyz|ab}/N) = \lim_N(V^B_{xyz|ab}/N)$.

$$\begin{aligned}
&\lim_N(\hat{V}_{xw,yz}/N) \\
&= \lim_N \frac{1}{N}\sum_k \frac{1}{N_k^2}\left\{X_{x|1}X_{w|1}X_{yz|2} + X_{xw|1}X_{y|2}X_{z|2} - X_{xw|1}X_{yz|2}\right\} \\
&= \lim_N \sum_k \frac{n_1 n_2}{N_k^2}\left\{\frac{n_1}{N}\frac{X_{x|1}}{n_1}\frac{X_{w|1}}{n_1}\frac{X_{yz|2}}{n_2} + \frac{n_2}{N}\frac{X_{xw|1}}{n_1}\frac{X_{y|2}}{n_2}\frac{X_{z|2}}{n_2} - \frac{1}{N}\frac{X_{xw|1}}{n_1}\frac{X_{yz|2}}{n_2}\right\}
\end{aligned}$$

$$= \sum_k \frac{\alpha_1\alpha_2}{(\sum_i \alpha_{ik})^2} \left\{\alpha_1\pi_{x|1}\pi_{w|1}\pi_{yz|2} + \alpha_2\pi_{xw|1}\pi_{y|2}\pi_{z|2} - 0\cdot\pi_{xw|1}\pi_{yz|2}\right\}$$

$$\begin{aligned}
&\lim_N(V_{xw,yz}/N)\\
&= \lim_N \frac{1}{N}\sum_k \frac{n_1n_2}{N_k^2}\left\{n_1'\pi_{x|1}\pi_{w|1}\pi_{yz|2} + n_2'\pi_{xw|1}\pi_{y|2}\pi_{z|2} + \pi_{xw|1}\pi_{yz|2}\right\}\\
&= \lim_N \sum_k \frac{n_1n_2}{N_k^2}\left\{\frac{n_1'}{N}\pi_{x|1}\pi_{w|1}\pi_{yz|2} + \frac{n_2'}{N}\pi_{xw|1}\pi_{y|2}\pi_{z|2} + \frac{1}{N}\pi_{xw|1}\pi_{yz|2}\right\}\\
&= \sum_k \frac{\alpha_1\alpha_2}{(\sum_i \alpha_{ik})^2}\left\{\alpha_1\pi_{x|1}\pi_{w|1}\pi_{yz|2} + \alpha_2\pi_{xw|1}\pi_{y|2}\pi_{z|2} + 0\cdot\pi_{xw|1}\pi_{yz|2}\right\}
\end{aligned}$$

Thus $\lim_N(\hat{V}_{xw,yz}/N) = \lim_N(V_{xw,yz}/N)$. We recall equation (3.12). Comparing (3.29) with (3.27) shows U_{xywz} is dually consistent, hence, also $\widehat{\text{Cov}}(\hat{\Psi}_{xy}, \hat{\Psi}_{xz})$. In contrast to U_{xywz}, U_{xyz} does not have exactly the same structure as (3.25). Note that $\hat{v}^A_{xyz|abk}$ and $\hat{v}^B_{xyz|abk}$ are symmetric in y and z. Also, note $v^B_{xyz|abk} = \Psi_{xy}v^B_{yxz|abk}$ and $v^B_{xyz|abk} = \Psi_{xz}v^B_{zxy|abk}$ for $a \neq b$. Hence $\frac{\lim_M V^B_{xyz|abk}}{\lim_M(C_{xy}/M)(C_{xz}/M)}$ cannot only be estimated by $\frac{\hat{V}^B_{xyz|abk}}{C_{xy}C_{xz}}$, but also by $\frac{\hat{V}^B_{yxz|abk}}{C_{yx}C_{xz}}$ and $\frac{\hat{V}^B_{yxz|abk}}{C_{xy}C_{zx}}$. The estimator U_{xyz} (3.28) was constructed by averaging over 3 such terms each. Therefore U_{xyz} does indeed converge to expression (3.25) under both limiting models. □

3.4 Generalised Variance and Covariance Estimators

From $\Psi_{xy}\Psi_{yz} = \Psi_{xz}$ follows $\log\Psi_{xy} + \log\Psi_{yz} = \log\Psi_{xz}$. Hence for $J \geq 3$, L_{xy} is not an unique estimator for $\log\Psi_{xy}$. Instead of estimating $\log\Psi_{xy}$ by L_{xy}, we use the generalised estimator

$$\bar{L}_{xy} := \frac{1}{J}\sum_{z=1}^{J}(L_{xz} - L_{yz}) = (L_{x+} - L_{y+})/J \tag{3.33}$$

as introduced by Greenland (1989) and originally suggested by Mickey and Elashoff (1985). Note $\bar{L}_{xy} = -\bar{L}_{yx}$ and $\bar{L}_{xz} = \bar{L}_{xy} + \bar{L}_{xz}$, but $L_{xy} = -L_{yx}$ and $L_{xz} \neq L_{xy} + L_{xz}$. $\bar{L}_{xy}$ is a linear combination of the L_{xy}'s, hence, variances and covariances of the generalised estimators can be easily computed from the variances and covariances of the L_{xy}'s. If any subscript of the U_{xywz}'s contains a "+" sign, then we sum over this index, e.g. $U_{x+xz} = \sum_h U_{xhxz}$.

Now we write

$$\begin{aligned}
&\text{Cov}(\bar{L}_{xy}, \bar{L}_{wz}) = \text{Cov}(1/J \sum_h L_{xh} - L_{yh}, 1/J \sum_i L_{wi} - L_{zi}) \\
&= \frac{1}{J^2} \sum_{h,i} \{\text{Cov}(L_{xh}, L_{wi}) + \text{Cov}(L_{yh}, L_{zi}) - \text{Cov}(L_{xh}, L_{zi}) - \text{Cov}(L_{yh}, L_{wi})\} \\
&= \frac{1}{J^2} \sum_{i} \{\text{Cov}(L_{xi}, L_{wi}) + \text{Cov}(L_{yi}, L_{zi}) - \text{Cov}(L_{xi}, L_{zi}) - \text{Cov}(L_{yi}, L_{wi})\} \\
&+ \frac{1}{J^2} \sum_{i \neq h} \{\text{Cov}(L_{xh}, L_{wi}) + \text{Cov}(L_{yh}, L_{zi}) - \text{Cov}(L_{xh}, L_{zi}) - \text{Cov}(L_{yh}, L_{wi})\}
\end{aligned} \tag{3.34}$$

and express $\sum_{h \neq i} \text{Cov}(L_{xh}, L_{wi})$ as

$$\begin{aligned}
\sum_{h \neq i} \text{Cov}(L_{xh}, L_{wi}) = &\sum_{\substack{h \\ (i = x)}} \text{Cov}(L_{xh}, L_{wx}) + \sum_{\substack{i \\ (h = w)}} \text{Cov}(L_{xw}, L_{wi}) \\
&- \text{Cov}(L_{xw}, L_{wx}) + \sum_{\substack{h, i \notin \{x, y\} \\ i \neq h}} \text{Cov}(L_{xh}, L_{wi}) \\
&= -\sum_i \text{Cov}(L_{xw}, L_{xi}) - \sum_i \text{Cov}(L_{wx}, L_{wi}) \qquad (3.35) \\
&+ \text{Cov}(L_{xw}, L_{xw}) + \sum_{\substack{h, i \notin \{x, y\} \\ h \neq i}} \text{Cov}(L_{xh}, L_{wi}) \qquad (3.36)
\end{aligned}$$

By combining (3.34) and (3.35), we estimate $\sum_{h,i} \text{Cov}(L_{xh}, L_{wi})$ by

$$U^+_{xw} := \sum_{h,i} \widehat{\text{Cov}}(L_{xh}, L_{wi}) = U_{+xw} - U_{xw+} - U_{wx+} + U_{xww} + S_{xw}$$

with $S_{xy} = \sum_{\substack{h,i \notin \{x,y\} \\ h \neq i}} U_{xhyi}$.
From $\sum_{h,i} \widehat{\text{Cov}}(L_{xh}, L_{xi}) = U_{x++}$ by definition, we can write

$$U^+_{xy} = \begin{cases} U^+_{xx} = U_{x++} = \sum_{h,i} U_{xhi} & , x = y \\ U^+_{xy} = U_{+xy} - U_{xy+} - U_{yx+} + U_{xy} + S_{xy} & , x \neq y \end{cases} \tag{3.37}$$

Summarising, we can express $\widehat{\text{Cov}}(\bar{L}_{xy}, \bar{L}_{wz})$ as

$$\bar{U}_{xywz} := \widehat{\text{Cov}}(\bar{L}_{xy}, \bar{L}_{wz}) = \frac{1}{J^2}\{U^+_{xw} - U^+_{xz} - U^+_{yw} + U^+_{yz}\} \tag{3.38}$$

with the special cases

$$\bar{U}_{xyz} := \widehat{\text{Cov}}(\bar{L}_{xy}, \bar{L}_{xz}) = \frac{1}{J^2}\{U_{x++} - U^+_{xz} - U^+_{yx} + U^+_{yz}\} (\equiv \frac{1}{J^2}\{U^+_{xx} - U^+_{xz} - U^+_{yx} + U^+_{yz}\})$$

and

$$\bar{U}_{xyy} := \widehat{\text{Var}}(\bar{L}_{xy}) = \frac{1}{J^2}\{U_{x++} - 2U^+_{xy} + U_{y++}\} (\equiv \frac{1}{J^2}\{U^+_{xx} - U^+_{xy} - U^+_{yx} + U^+_{yy}\}).$$

Formula (3.38) is generally applicable for all indices x, y, w, z and is identical to Greenland's formula. However, Greenland defined the term U^+_{xy} ($x \neq y$) as $U^+_{xy} := U_{+xy} - U_{xy+} - U_{xy+} + U_{xyy}$, which does not contain S_{xy}, because $\text{Cov}(L_{xy}, L_{wz}) = 0$ for the binomial and multinomial sampling scheme. Equation (3.26) shows that this is generally not true for our sampling scheme of two rows of multiple responses, and therefore the term S is indeed required.

3.5 Extended Generalised Estimators

Suppose $r > 2$. Each odds ratio $\Psi_{xy|ab}$ can be computed from the set of $(r-1) \times (J-1)$ local odds ratios $\{\Psi_{i,i+1|j,j+1}; i = 1, \ldots, J; j = 1, \ldots, r\}$ by

$$\Psi_{xy|ab} = \prod_{i=x}^{y} \prod_{j=a}^{b} \Psi_{i,i+1|j,j+1},$$

which follows from $\Psi_{xy|ab}\Psi_{yz|ab} = \Psi_{xz|ab}$ and $\Psi_{xy|ab}\Psi_{xy|bc} = \Psi_{xy|ac}$. In a similar way, we can compute $\Psi_{xy|ab}$ by

$$\Psi_{xy|ab} = \prod_{i=1}^{J} \frac{\Psi_{xi|ab}}{\Psi_{yi|ab}} = \prod_{j=1}^{r} \frac{\Psi_{xy|aj}}{\Psi_{xy|bj}} = \prod_{i=1}^{J} \prod_{j=1}^{r} \frac{\Psi_{xi|aj}}{\Psi_{xi|bj}} \frac{\Psi_{yi|aj}}{\Psi_{yi|bj}}$$

leading to the generalised estimator $\bar{L}_{xy|ab}$ estimating $\log \Psi_{xy|ab}$ by

$$\bar{L}_{xy|ab} = \frac{1}{rJ} \sum_{i=1}^{J} \sum_{j=1}^{r} L_{xi|aj} - L_{xi|bj} - L_{yi|aj} + L_{yi|bj} = \frac{1}{rJ}(L_{x+|a+} - L_{x+|b+} - L_{y+|a+} + L_{y+|b+}).$$

Now we could proceed with deriving a generalised (co)variance estimator for $\text{Cov}(\bar{L}_{xy|ab}, \bar{L}_{wz|ac})$, but this requires an estimator for $\text{Cov}(L_{xy|ab}, L_{wz|ac})$. It actually requires several estimators, because we also need to consider special cases such as $x = w$ or $b = c$. We leave the derivation of such estimators and a generalised (co)variance estimator for future research.

3.6 Example

We reconsider the UTI data in Table 1.1 on page 2 with items: A-oral, B-condom, C-lubricated condom, D-spermicide, and E-diaphragm. For simplicity, we exclude item E due to zero cell counts and therefore avoid amending the MH esti-

mator and its variance estimator. The MH approach gives $\{\bar{L}_{AB}, \bar{L}_{AC}, \bar{L}_{AD}, \bar{L}_{BC}, \bar{L}_{BD}, \bar{L}_{CD}\}$=$\{0.28\ , -0.43, -0.45, -0.70, -0.73, -0.02\}$ with standard errors $\{0.21, 0.25, 0.29, 0.13, 0.20, 0.21\}$ by applying formula (3.38). The bootstrap and the generalised (co)variance estimates with $B = 50,000$ can be found in Table 3.1. For instance, comparing the UTI effects for the contraceptives "oral" and "lubricated condom" the MH estimator is -0.43 with s.e. 0.25, which gives the 95% confidence interval $(-0.92, +0.06)$. The odds of using "oral" (rather than using "lubricated condom") for a woman without UTI history are $\exp(-0.43) = 0.65$ times the odds of using "oral" (rather than using "lubricated condom") for a woman with UTI history.

Table 3.1: The "bootstrap" with $B = 50,000$ and generalised ($\bar{U}$) (co)variance estimates of $\{\bar{L}_{ij},\ i,j = A,\ldots,D\}$, $100\times$ (co)variance

	$\bar{L}_{AB}$	$\bar{L}_{AC}$	$\bar{L}_{AD}$	$\bar{L}_{BC}$	$\bar{L}_{BD}$	$\bar{L}_{CD}$
$\bar{L}_{AB}$	4.39 (4.43)	4.50 (4.34)	4.50 (4.43)	0.11 (-0.08)	0.11 (0.00)	0.00 (0.66)
$\bar{L}_{AC}$		6.55 (6.19)	5.47 (5.01)	2.05 (1.84)	0.97 (0.01)	-1.08 (-1.18)
$\bar{L}_{AD}$			9.07 (8.32)	0.97 (0.08)	4.58 (3.90)	3.61 (3.32)
$\bar{L}_{BC}$				1.94 (1.81)	0.87 (0.80)	-1.08 (-1.04)
$\bar{L}_{BD}$					4.47 (4.06)	3.61 (3.18)
$\bar{L}_{CD}$						4.69 (4.34)

The odds ratio $\Psi_{xy|ab}$ allows us to describe the relationship of the odds between two items. In contrast, the odds ratio Ψ^{x}_{ab} only describes the odds ratio based on the item x, chosen versus not chosen.

3.7 Simulation Study

3.7.1 Simulation Scheme

In this section, we conduct a simulation study to investigate the performance of the proposed estimators U_{xyy} and U_{xyz}, and the generalised estimators $\bar{U}_{xyy}$ and $\bar{U}_{xyz}$. We also want to double check the correctness of the proposed dually consistent co- and variance estimators. Unfortunately, we are not aware of any model-based approach to estimate $\Psi_{xy|ab}$ or $\log \Psi_{xy|ab}$. To investigate the performance of the generalised estimator, we choose $J = 3$, since for $J = 3$, estimators $\bar{U}_{xyy}$ and U_{xyy} differ generally, but for $J = 2$ they are identical. A disadvantage of $J = 3$ is that we cannot investigate U_{xywz}.

For given Ψ_{xy}, we fix the marginal probabilities of the first row by setting $\pi_{x|1k} = 0.50$ for all $x = 1, \ldots, J$. Then we set $\pi_{1|2k} = 1/(1 + \Psi_{xy})$ and $\pi_{x|2k} = \frac{\Psi_{xy}}{1+\Psi_{xy}}$ for $x = 2, \ldots, J$. This ensures that the probabilities of the second row are balanced around $1/2$, for example $\Psi_{12} = 1$ gives $\pi_{1|2k} = \pi_{2|2k} = 1/2$. For simplicity, we also set $\Psi = \Psi_{12} = \Psi_{13}$ and $N_k = N_1 = \cdots = N_K$.

Let Y_x indicate whether a subject selects item x. Given row a and stratum k, if a subject selects item x, then $Y_x = 1$; otherwise, $Y_x = 0$. Again as in the previous chapter (see Section 2.5 on page 59), we define the pairwise dependency between items x and y in form of an odds ratio $\theta_{xy|ak}$ as

$$\theta_{xy|ik} = \frac{P(Y_x = 1, Y_y = 1|ak)P(Y_x = 0, Y_y = 0|ak)}{P(Y_x = 0, Y_y = 1|ak)P(Y_x = 1, Y_y = 0|ak)}.$$

For convenience, we assume a constant association $\theta = \theta_{xy|ik}$ for all items $x, y = 1, \ldots, J$, rows $a = 1, 2$ and strata $k = 1, \ldots, K$.

When we fix the covariance between two items by $\text{Cov}(Y_x, Y_y) = -\pi_{x|ak}\pi_{y|ak}$,

then $\pi_{xy|ak} = P(Y_x = 1, Y_y = 1|ak) = 0$ and consequently $\theta_{xy|ak} = 0$. Fixing the covariance in such a way for all pairs of items yields the multinomial distribution (see Appendix B on page 296). In the the following, when we set $\theta = \theta_{xy|ak} = 0$, we sample from the multinomial distribution with probability $\pi_{x|ak}$ of choosing the xth item for row $a = 1, \ldots, r$ and stratum $k = 1, \ldots, K$. For the multinomial distribution $\sum_{x=1}^{J} \pi_{xa|k} \leq 1$, meaning the above settings for $\pi_{x|ak}$ are invalid. Instead we set $\pi_{x|1k} = 1/(J+1)$, $\pi_{1|2k} = 1/[(J-1)\Psi + 1]$ and $\pi_{x|2k} = \Psi\pi_{1|2k}$ for $x \geq 2$.

The number of bootstrap simulations was chosen as $B = 400$ and the number of simulated datasets as $20,000$. We record the mean and m.s.e. (mean squared error) of the bootstrap estimate of (co-)variance (denoted by Var^{BT} and Cov^{BT}), of the newly proposed (co)variance estimators [U_{xyy} defined by (3.20) on page 99, U_{xyz} by (3.28) on page 105, and $\bar{U}$ defined by (3.38) on page 111] and also of the "old" (co)variance estimators proposed by Greenland (1989) based on multinomial sampling [estimator U^{old}, see page 105, and estimator $\bar{U}^{old}$ also using equation (3.38) on page 111 but replacing U by U^{old} and deleting S_{xy}]. The empirical variance (denoted by Var^{emp}) and covariance (denoted by Cov^{emp}) of the L's and $\bar{L}$'s over all simulations are regarded as the true (co-)variances. The number of simulations n_{MH} for which the MH estimators were undefined is also recorded.

3.7.2 Simulation Results

First we compare how the MH and the generalised MH estimator perform for various configurations. Ideally, we want an estimator with no or small bias (difference between empirical mean and true value of parameter) and with low (empirical) variance. The mean squared error (m.s.e) summarizes both criteria, since m.s.e = bias^2 + variance.

Table 3.2: Simulation results for log odds ratio estimators L and $\bar{L}$

K, N_k, Ψ, θ,n_{MH}	mean L_{12}, L_{13}	mean $\bar{L}_{12}$,$\bar{L}_{13}$	m.s.e. L_{12}, L_{13}	m.s.e. $\bar{L}_{12}$,$\bar{L}_{13}$
5 , 20 , 4 , 0 ,101	$1.489, 1.481$	$1.487, 1.484$	$0.578, 0.569$	$0.526, 0.525$
5 , 20 , 4 , 1 ,1	$1.438, 1.438$	$1.437, 1.439$	$0.165, 0.166$	$0.161, 0.162$
5 , 20 , 4 , 10 ,1	$1.431, 1.434$	$1.431, 1.433$	$0.130, 0.134$	$0.128, 0.131$
20 , 5 , 4 , 0 ,1933	$1.333, 1.334$	$1.335, 1.332$	$0.629, 0.619$	$0.482, 0.484$
20 , 5 , 4 , 1 ,22	$1.464, 1.465$	$1.464, 1.465$	$0.265, 0.265$	$0.237, 0.238$
20 , 5 , 4 , 10 , 18	$1.460, 1.463$	$1.461, 1.462$	$0.215, 0.220$	$0.203, 0.204$
1 , 500 , 4 , 0 ,0	$1.397, 1.401$	$1.397, 1.401$	$1.054, 1.060$	$1.054, 1.060$
1 , 500 , 4 , 1 ,0	$1.395, 1.394$	$1.395, 1.394$	$1.002, 1.000$	$1.002, 1.000$
1 , 500 , 4 , 4 ,0	$1.393, 1.393$	$1.393, 1.393$	$0.992, 0.991$	$0.992, 0.991$
100 , 5 , 4 , 0 ,1	$1.427, 1.427$	$1.427, 1.428$	$1.187, 1.190$	$1.145, 1.144$
100 , 5 , 4 , 1 ,0	$1.399, 1.397$	$1.398, 1.398$	$1.021, 1.017$	$1.016, 1.014$
100 , 5 , 4 , 4 ,0	$1.401, 1.402$	$1.401, 1.402$	$1.016, 1.017$	$1.013, 1.015$

log(4)= 1.386294

Table 3.2 shows the generalised estimator $\bar{L}$ performs slightly better than L for the sparse data case ($N_k = 5$ and $K = 20, 100$, $N_k = 20$ and $K = 5$), but it seems the larger K is, the smaller the difference is between $\bar{L}$ and L. For the large stratum case ($N_k = 500$ and $K = 1$), both estimators perform equally well.

Next we consider the performance of the existing and newly proposed (co)variance estimators. Ideally the mean of the formula variance (U's) should equal the empirical variance of the MH estimator (likewise for the bootstrap), which would indicate no bias, and the variance (or the combined measure m.s.e.) should be as low as possible.

Table 3.3 shows the simulation results of the variance estimators for various scenarios. The newly proposed estimators, U_{122} and U_{123}, perform better than the bootstrap estimates of (co-)variance except for $K = 20$ and $N_k = 5$. They are also superior to U_{122}^{old} and U_{123}^{old} for $\theta > 0$. Only for $\theta = 0$, U and U^{old} are identical, because $U^{add} = 0$ for $\theta = 0$ due to the impossible event of observing the pairwise

Table 3.3: Simulation results for the variance and covariance estimators of the log odds ratio estimators

	$\text{Var}^{emp}(L_{12}), \text{Cov}^{emp}(L_{12}, L_{13})$		$\text{Var}^{emp}(\bar{L}_{12}), \text{Cov}^{emp}(\bar{L}_{12}, \bar{L}_{13})$	
	$\text{Var}^{BT}(L_{12}), \text{Cov}^{BT}(L_{12}, L_{13})$		$\text{Var}^{BT}(\bar{L}_{12}), \text{Cov}^{BT}(\bar{L}_{12}, \bar{L}_{13})$	
	U_{122}, U_{123}		$\bar{U}_{122}, \bar{U}_{123}$	
K, N_k, Ψ, θ	$U^{old}_{122}, U^{old}_{123}$		$\bar{U}^{old}_{122}, \bar{U}^{old}_{123}$	
n_{MH}	100×mean	10000×mse	100×mean	100000×mse
5 , 20 , 4 , 1	16.25, 12.50	−, −	15.82, 12.95	−, −
1	20.38, 13.95	97.23, 39.79	18.84, 13.95	65.99, 45.07
	15.39, 11.71	37.39, 30.74	14.96, 12.13	33.80, 31.98
	24.54, 16.13	107.3, 41.56	25.40, 12.88	144.9, 22.13
5 , 20 , 4 , 10	12.78, 11.13	−, −	12.61, 11.33	−, −
1	16.73, 13.52	79.41, 45.06	16.00, 13.52	65.34, 52.23
	12.07, 10.21	34.00, 32.75	11.85, 10.43	33.22, 32.65
	24.83, 16.64	186.01, 61.60	25.81, 13.25	226.1, 27.22
20 , 5 , 4 , 1	25.94, 16.14	−, −	23.13, 18.88	−, −
22	29.71, 12.74	144.7, 151.6	23.73, 12.74	44.24, 42.80
	23.76, 14.75	198.0, 84.31	21.28, 17.23	121.3, 103.6
	37.58, 20.19	394.5, 88.70	35.82, 19.02	405.2, 79.28
20 , 5 , 4 , 10	20.926, 16.166	−, −	19.718, 17.516	−, −
18	25.54, 16.52	110.0, 53.80	22.47, 16.52	57.50, 41.16
	19.38, 14.83	177.5, 116.3	18.19, 16.06	137.5, 127.6
	39.51, 23.50	626.5, 165.3	38.80, 21.05	642.2, 120.7
1 , 500 , 4 , 0	7.796, 5.286	−, −	7.796, 5.286	−, −
0	8.194, 5.616	3.966, 4.230	8.194, 5.616	3.966, 4.230
	7.872, 5.345	0.539, 0.552	7.872, 5.345	0.539, 0.552
	7.872, 5.345	0.539, 0.552	8.321, 4.027	1.304, 2.031
1 , 500 , 4 , 1	2.594, 2.084	−, −	2.594, 2.084	−, −
0	2.611, 2.100	0.218, 0.272	2.611, 2.100	0.218, 0.272
	2.542, 2.034	0.088, 0.084	2.540, 2.036	0.088, 0.084
	4.149, 2.843	2.499, 0.654	4.376, 2.179	3.299, 0.069
1 , 500 , 4 , 4	2.086, 1.757	−, −	2.086, 1.757	−, −
0	2.202, 1.853	0.191, 0.150	2.202, 1.853	0.191, 0.150
	2.138, 1.790	0.081, 0.076	2.136, 1.792	0.081, 0.076
	4.146, 2.839	4.325, 1.248	4.371, 2.178	5.339, 0.234
100 , 5 , 4 , 0	16.63, 4.525	−, −	12.68, 8.687	−, −
1	25.921, 0.445	267.0, 187.9	15.57, 0.445	35.93, 12.01
	15.982, 4.232	20.74, 1.152	12.037, 8.181	6.947, 4.353
	15.982, 4.232	20.74, 1.152	12.425, 7.116	9.569, 5.853
100 , 5 , 4 , 1	4.210, 2.612	−, −	3.798, 3.005	−, −
0	4.938, 2.336	3.514, 2.605	4.050, 2.336	1.073, 0.728
	4.087, 2.564	0.581, 0.284	3.699, 2.947	0.390, 0.324
	6.734, 3.576	7.085, 1.126	6.385, 3.375	7.324, 0.376
100 , 5 , 4 , 4	3.672, 2.613	−, −	3.415, 2.872	−, −
0	4.450, 2.680	2.795, 1.123	3.871, 2.680	1.258, 0.711
	3.565, 2.535	0.544, 0.336	3.316, 2.785	0.422, 0.372
	7.012, 3.976	11.96, 2.134	6.775, 3.619	12.02, 0.860

observation $(1, 1)$.

For example: For $K = 1$ and $N_k = 500$, U^{old} and U are identical for $\theta = 0$ and show almost no bias and have a low m.s.e., whereas the bootstrap method has a slighly larger bias and a much higher m.s.e. For the same setting, but for $\theta = 4$, U and the bootstrap method behave quite similarly, however U^{old} is now severely biased and also has a much higher m.s.e. than before.

When taking a closer look at Table 3.3, we see the bigger θ becomes, the bigger the difference between U^{old}_{xyz} and U_{xyz} becomes. We also see that $\bar{U}_{122}$ performs better than U_{122}, however, $\bar{U}_{123}$ does not perform better than U_{123}. This can be explained by noting that the m.s.e of U_{122} is higher than that of U_{123}. Now the $\bar{U}$'s are a linear combination of the U's, hence the m.s.e of the $\bar{U}$'s lies in between the mse's of U_{122} and U_{123}.

Generally, U^{old} cannot be recommended for $\theta > 0$, because the U^{old}'s are severely biased. For $\theta = 0$, the old and new estimators are identical. Overall the newly proposed (co-)variance estimators U_{xyy}, U_{xyz} and their generalised versions $\bar{U}_{xyy}$, $\bar{U}_{xyz}$ perform very well and can be highly recommended, if one is interested in the conditional association for two independent rows of multiple responses. Only for very sparse data and a small number of strata (e.g. $N_k \leq 20$ and $K \leq 5$) do we prefer the bootstrap estimator of (co)variance. Also, we expect the generalised estimators ($\bar{L}$ and $\bar{U}$) to perform even better for $J > 4$. We assume that U_{xywz} behaves similarly to U_{xyz} and U_{xyy}, due to the similar construction of the estimator.

The simulation results also confirm the correctness the proposed estimators. If they were incorrect, their performance would deteriorate for growing K or growing N_k, but the opposite is true. Their performance relative to the other estimators improves consistently for growing K and N_k when N_k and K, respectively,

remain fixed.

Chapter 4

Mantel-Haenszel Estimators for One Row of Multiple Responses per Stratum

4.1 Introduction

In this chapter, we consider the odds ratio estimation for one row of multiple responses with J outcome categories per stratum, forming K $2 \times J$ tables. The first row of each of the K tables comprises the positive responses and the second row the negative responses. For example, for the UTI data in Table 1.1 on page 2, we obtain one row of multiple responses per stratum merging women with and without a prior UTI history. Then we regard the positive and negative responses as two rows of a table forming, for each of the $K = 2$ strata, a $2 \times J = 2 \times 5$ table. The kth odds ratio is defined as

$$\Psi_{xy|k} = \frac{\pi_{x|k}\bar{\pi}_{y|k}}{\pi_{y|k}\bar{\pi}_{x|k}}, \tag{4.1}$$

where $\pi_{x|k}$ denotes the probability of a positive response and $\bar{\pi}_{x|k}$ that of a negative response for item $x = 1 \ldots, J$ and stratum $k = 1, \ldots, K$. We use the same notations as in Chapter 2, for instance $X_{x|k}$ and $\bar{X}_{x|k}$ denote the corresponding observations. In contrast to Chapters 2 and 3, we omit subscript a referring to the ath row, because there is only one. This sampling scheme can also be regarded as J dependent binomials for each of the K tables. Due to the dependence between items, the ordinary MH estimator is not dually consistent anymore, but only consistent under the large stratum limiting model (model I), as noted by Yanagawa and Fujii (1995).

First we propose a model-based estimator in Section 4.2. Then in Section 4.3, we show that the ordinary MH estimator is not consistent under model II, but still consistent under model I. In Section 4.4, we propose a new MH estimator that is dually consistent. For this new MH-type estimator, we derive in Section 4.5 the asymptotic variance for both limiting models and derive in Section 4.6 a dually consistent variance estimator. Section 4.7 illustrates the method on the UTI data. Then we conduct a simulation study in Section 4.8 investigating the performance of the ordinary and new MH-type estimator and their variance estimators.

4.2 An Odds Ratio Estimator

Let $\mathbf{Y} = (Y_1, \ldots, Y_J)$ denote the multiple response variable with J items, that is, having J outcome categories and allowing multiple choices. Let W denote a control variable having K categories. Then $\pi_{x|k}$ is the probability that level x of $\mathbf{Y}$ is chosen, when the control variable is at level $k \in \{1, \ldots, K\}$. We consider a

logistic regression model having the form

$$\log\left(\frac{\pi_{x|k}}{1-\pi_{x|k}}\right) = \tau_k + \beta_x. \tag{4.2}$$

The odds ratio for the xth and yth outcome of variable $\mathbf{Y}$ is

$$\exp(\beta_x - \beta_y) = \frac{\pi_{x|k}(1-\pi_{y|k})}{(1-\pi_{x|k})\pi_{y|k}} =: \Psi_{xy}. \tag{4.3}$$

The interpretation of Ψ_{xy} is that the odds of making a positive response at level x of $\mathbf{Y}$ are $\exp(\beta_x - \beta_y)$ times the odds of making a positive response at level y of $\mathbf{Y}$ independently of the level of W. Estimates can be be obtained from generalised estimation equations (GEE), see Section 5.2.2 on page 150, or from ML estimation, incorporating the correlation between categories of $\mathbf{Y}$. However, GEE and ML estimation may not be consistent under model II, because model parameters τ_k and sample size grow simultaneously. Alternatively, we may use the common MH estimator.

4.3 The Ordinary Mantel-Haenszel Estimator

Note $N_k = n_{1k} =: n_k$. Again, we assume a common odds ratio

$$\Psi_{xy} = \Psi_{xy|1} = \ldots \Psi_{xy|K} \tag{4.4}$$

from which follows

$$\pi_{x|k}\bar{\pi}_{y|k} = \Psi_{xy}\pi_{y|k}\bar{\pi}_{x|k}. \tag{4.5}$$

The ordinary Mantel-Haenszel estimator has the form

$$\hat{\Psi}_{xy} = C_{xy}/C_{yx} \tag{4.6}$$

with $C_{xy} = \sum_{k=1}^{K} c_{xy|k}$ and $c_{xy|k} = X_{x|1k}\bar{X}_{y|2k}/n_k$.

Theorem 4.3.1. *The Mantel-Haenszel estimator is not consistent under model II, but still consistent under model I, as mentioned by Yanagawa and Fujii (1995).*

Proof. Sparse-Data: Under model II we can write

$$\begin{aligned}
\hat{\Psi}_{xy} - \Psi_{xy} &= \frac{C_{xy} - \Psi_{xy}C_{yx}}{C_{yx}} = \frac{\sum_{k=1}^{K} c_{xy|k} - \Psi_{xy}c_{yx|k}}{\sum_{k=1}^{K} c_{yx|k}} \\
&= \frac{(\sum_{k=1}^{K} c_{xy|k} - \Psi_{xy}c_{yx|k})/K}{\sum_{k=1}^{K} c_{yx|k}/K} = \frac{(C_{xy} - \Psi_{xy}C_{yx})/K}{C_{yx}/K} \\
&= \frac{\sum_{k=1}^{K} \omega_{xy|k}/K}{\sum_{k=1}^{K} c_{yx|k}/K} = \frac{\Omega_{xy}/K}{C_{yx}/K}.
\end{aligned} \tag{4.7}$$

with $\omega_{xy|k} = c_{xy|k} - \Psi_{xy}c_{yx|k}$ and $\Omega = \sum_k \omega_k$.

The term $c_{xy|k}$ is a bounded random variable under limiting model II, hence, the variance of C_{xy} is $o(K^2)$ and Theorem 2.8.2 on page 74 implies $\frac{1}{K}(\Omega_{xy} - \mathbb{E}\Omega_{xy})$ $\rightarrow_p 0$. We have

$$\begin{aligned}
\mathbb{E}c_{xy|k} &= \frac{1}{n_k}\mathbb{E}X_{x|k}\bar{X}_{y|k} = \frac{1}{n_k}\mathbb{E}X_{x|k}(n_k - X_{y|k}) = \frac{1}{n_k}\left(n_k\mathbb{E}X_{x|k} - \mathbb{E}X_{x|k}X_{y|k}\right) \\
&= \frac{1}{n_k}\left(n_k^2\pi_{x|k} - n_k\left[n_k'\pi_{x|k}\pi_{y|k} + \pi_{xy|k}\right]\right) \\
&= n_k\pi_{x|k} - n_k\pi_{x|k}\pi_{y|k} + \pi_{x|k}\pi_{y|k} - \pi_{xy|k} \\
&= n_k\pi_{x|k}\bar{\pi}_{y|k} + (\pi_{x|k}\pi_{y|k} - \pi_{xy|k})
\end{aligned} \tag{4.8}$$

by using

$$\mathbb{E}X_{x|k}X_{y|k} = n_k[n_k'\pi_{x|k}\pi_{y|k} + \pi_{xy|k}^{11}] \tag{4.9}$$

from (2.20) on page 74 with $n'_k = n_k - 1$. Thus

$$\begin{aligned}
&\mathbb{E}\Omega_{xy} = \mathbb{E}(C_{xy} - \Psi C_{yx}) \\
&= \sum_k \left\{ n_k \pi_{x|k} \bar{\pi}_{y|k} + (\pi_{x|k}\pi_{y|k} - \pi_{xy|k}) - \Psi_{xy} \left(n_k \pi_{y|k} \bar{\pi}_{x|k} + (\pi_{y|k}\pi_{x|k} - \pi_{xy|k}) \right) \right\} \\
&= \sum_k \left\{ n_k \pi_{x|k} \bar{\pi}_{y|k} - n_k \pi_{x|k} \bar{\pi}_{y|k} + (\pi_{x|k}\pi_{y|k} - \pi_{xy|k})(1 - \Psi_{xy}) \right\} \\
&= (1 - \Psi_{xy}) \sum_k (\pi_{x|k}\pi_{y|k} - \pi_{xy|k}). \qquad (4.10)
\end{aligned}$$

Assuming independence between items x and y, we have $\pi_{xy|k} = \pi_{x|k}\pi_{y|k}$ and consequently $\mathbb{E}\Omega_{xy} = \mathbb{E}(C_{xy} - \Psi C_{yx}) = 0$. We also find $\mathbb{E}(C_{xy} - \Psi C_{yx}) = 0$ for $\Psi_{xy} = 1$. However, in general $\mathbb{E}\Omega_{xy} \neq 0$, hence $\frac{1}{K}\Omega_{xy} \rightarrow_p c \neq 0$. That is, the numerator of (4.7) does not converge to zero. By applying the Chebyshev weak law of large numbers to the denominator, we have

$$\sum_{k=1}^{K} c_{xy|k}/K \overset{K\rightarrow\infty}{\longrightarrow}_p \lim_{K\rightarrow\infty} \sum_{k=1}^{K} \mathbb{E}(c_{xy|k})/K. \qquad (4.11)$$

This limit is finite and nonzero. We conclude from Slutsky's theorem that $\hat{\Psi}_{xy} - \Psi_{xy} \rightarrow_p c \neq 0$. Consequently, $\hat{\Psi}_{xy}$ is not consistent under model II.

Large-Stratum: Now we consider the case $N \rightarrow \infty$ with $N\alpha_k = n_k$ and $0 < \alpha_k < 1$, that is, as N approaches infinity so the number of subjects n_k for all strata k also approaches infinity.

Now

$$\begin{aligned}
\hat{\Psi}_{xy} &= \frac{\sum_{k=1}^{K} X_{x|k}\bar{X}_{y|k}/n_k}{\sum_{k=1}^{K} X_{y|k}\bar{X}_{x|k}/n_k} = \frac{\sum_k \frac{1}{n_k N} X_{x|k}\bar{X}_{y|k}}{\sum_k \frac{1}{n_k N} X_{y|k}\bar{X}_{x|k}} = \frac{\sum_k \frac{n_k}{N}\frac{X_{x|k}}{n_k}\frac{\bar{X}_{y|k}}{n_k}}{\sum_k \frac{n_k}{N}\frac{X_{y|k}}{n_k}\frac{\bar{X}_{x|k}}{n_k}} \\
&\overset{N\rightarrow\infty}{\longrightarrow}_p \frac{\sum_k \alpha_k \pi_{x|k}\bar{\pi}_{y|k}}{\sum_k \alpha_k \pi_{y|k}\bar{\pi}_{x|k}} = \Psi_{xy}\frac{\sum_k \alpha_k \pi_{y|k}\bar{\pi}_{x|k}}{\sum_k \alpha_k \pi_{y|k}\bar{\pi}_{x|k}} = \Psi_{xy}, \qquad (4.12)
\end{aligned}$$

by (4.5), that is the consistency under model I. □

Remark 4.3.2. We showed that the ordinary MH estimator is not only dually consistent under independence of items, but also when $\Psi_{xy} = 1$.

4.4 A New Mantel-Haenszel Type Estimator $\tilde{\Psi}$

We propose the following new estimator for the common odds ratio Ψ_{xy}

$$\tilde{\Psi}_{xy} = \frac{\tilde{C}_{xy}}{\tilde{C}_{yx}} \tag{4.13}$$

with $\tilde{C}_{xy} = \sum_k \tilde{c}_{xy|k}$ and $\tilde{c}_{xy|k} = (X_{x|k}\bar{X}_{y|k} - X^{10}_{xy|k})/n'_k$; where by definition $X^{10}_{xy|k} = X^{01}_{yx|k}$. Note, $\tilde{c}_{xy|k}$ differs from $c_{xy|k}$ only by the extra term $X^{10}_{xy|k}$ and n_k is replaced by $n'_k := n_k - 1$. Under independence of items x and y, we have $\mathbb{E}X^{10}_{xy|k} = \pi^{10}_{xy|k} = \pi_{x|k}\bar{\pi}_{y|k} = \mathbb{E}X_{x|k}\mathbb{E}\bar{X}_{y|k}/n_k$, hence $\mathbb{E}\tilde{c}_{xy|k} = (\mathbb{E}X_{x|k}\mathbb{E}\bar{X}_{y|k} - 1/n_k\mathbb{E}X_{x|k}\mathbb{E}\bar{X}_{y|k})/n'_k = \mathbb{E}X_{x|k}\mathbb{E}\bar{X}_{y|k}/n_k = \mathbb{E}c_{xy|k}$. We conclude the construction of $\tilde{c}_{xy|k}$ is consistent with $c_{xy|k}$ when items are independent.

Theorem 4.4.1. *The new estimator $\tilde{\Psi}_{xy}$ is dually consistent.*

Proof. Sparse Data:

Similarly as before, we can write

$$\begin{aligned}\tilde{\Psi}_{xy} - \Psi_{xy} &= \frac{(\sum_{k=1}^K \tilde{c}_{xy|k} - \Psi_{xy}\tilde{c}_{yx|k})/K}{\sum_{k=1}^K \tilde{c}_{yx|k}/K} = \frac{(\tilde{C}_{xy} - \Psi_{xy}\tilde{C}_{yx})/K}{\tilde{C}_{yx}/K} \\ &= \frac{(\sum_{k=1}^K \tilde{\omega}_{xy|k})/K}{\sum_{k=1}^K \tilde{c}_{yx|k}/K} = \frac{\tilde{\Omega}_{xy}/K}{\tilde{C}_{yx}/K}\end{aligned} \tag{4.14}$$

with $\tilde{\omega}_{xy|k} = \tilde{c}_{xy|k} - \Psi_{xy}\tilde{c}_{yx|k}$ and $\tilde{\Omega} = \sum_k \tilde{\omega}_k$. We have

$$\mathbb{E}\tilde{c}_{xy|k} = \mathbb{E}(X_{x|k}\bar{X}_{y|k} - X^{10}_{xy|k})/n'_k = \frac{1}{n'_k}\left(\mathbb{E}X_{x|k}(n_k - X_{y|k}) - \mathbb{E}X^{10}_{xy|k}\right)$$

$$
\begin{aligned}
&= \frac{1}{n'_k}\left(n_k \mathbb{E}X_{x|k} - \mathbb{E}X_{x|k}X_{y|k} - \mathbb{E}X^{10}_{xy|k}\right) \\
&= \frac{1}{n'_k}\left(n_k^2\pi_{x|k} - n_k(n'_k\pi_{x|k}\pi_{y|k} + \pi^{11}_{xy|k}) - n_k\pi^{10}_{xy|k}\right) \\
&= \frac{1}{n'_k}\left(n_k n'_k(\pi_{x|k} - \pi_{x|k}\pi_{y|k}) + n_k(\pi_{x|k} - \pi^{11}_{xy|k} - \pi^{10}_{xy|k})\right) \\
&= n_k\pi_{x|k}\bar{\pi}_{y|k} + \frac{n_k}{n'_k}(\pi_{x|k} - \pi_{x|k}) = n_k\pi_{x|k}\bar{\pi}_{y|k}, \qquad (4.15)
\end{aligned}
$$

hence, $\mathbb{E}\tilde{\Omega}_{xy} = \mathbb{E}(\tilde{C}_{xy} - \Psi_{xy}\tilde{C}_{yx}) = 0$. Under independence of items, we also have $\mathbb{E}c_{xy|k} = n_k\pi_{x|k}\bar{\pi}_{y|k}$. We apply Chebyshev's weak law of large numbers and find

$$
\tilde{C}_{xy}/K = \sum_{k=1}^{K}\tilde{c}_{xy|k}/K \overset{K\to\infty}{\longrightarrow}_p \lim_{K\to\infty}\sum_{k=1}^{K}\mathbb{E}(\tilde{c}_{xy|k})/K = \lim_{K\to\infty}\mathbb{E}\tilde{C}_{xy}/K. \qquad (4.16)
$$

It follows now from equation (4.14) and from applying Chebyshev's weak law of large numbers to the numerator together with $\mathbb{E}\tilde{\Omega}_{xy} = 0$ and equation (4.16) that the new estimator $\tilde{\Psi}_{xy}$ is consistent under model II, in contrast to $\hat{\Psi}_{xy}$.

Large Stratum: Again, we consider the term $\tilde{C}_{xy}$. $\mathbb{E}\tilde{c}_{xy} = n_k\pi_{x|k}\bar{\pi}_{y|k}$ by (4.15) , hence,

$$
\mathbb{E}\tilde{C}_{xy|k}/N = \sum_{k=1}^{K}\mathbb{E}\tilde{c}_{xy|k}/N = \sum_{k=1}^{K}\frac{n_k}{N}\pi_{x|k}\bar{\pi}_{y|k} = \sum_{k=1}^{K}\frac{n_k}{N}\pi_{x|k}\bar{\pi}_{y|k} \overset{N\to\infty}{\longrightarrow}_p \sum_{k=1}^{K}\alpha_k\pi_{x|k}\bar{\pi}_{y|k}. \qquad (4.17)
$$

Also

$$
\begin{aligned}
\tilde{C}_{xy|k}/N = \sum_{k=1}^{K}\tilde{c}_{xy|k}/N &= \sum_{k=1}^{K}(X_{x|k}\bar{X}_{y|k} - X^{10}_{xy|k})/(n'_k N) \\
&= \sum_{k=1}^{K}\frac{n_k^2}{n'_k N}\frac{X_{x|k}}{n_k}\frac{\bar{X}_{y|k}}{n_k} - \frac{n_k}{n'_k N}\frac{X^{10}_{xy|k}}{n_k} \\
&\overset{N\to\infty}{\longrightarrow}_p \sum_{k=1}^{K}\alpha_k\pi_{x|k}\bar{\pi}_{y|k} - 0\cdot\pi^{10}_{xy|k} = \sum_{k=1}^{K}\alpha_k\pi_{x|k}\bar{\pi}_{y|k}. \qquad (4.18)
\end{aligned}
$$

Thus, we have for both models

$$\tilde{C}_{xy|k}/M \overset{M\to\infty}{\longrightarrow}_p \lim_M \mathbb{E}\tilde{C}_{xy|k}/M \text{ with } M \in \{K, N\}. \tag{4.19}$$

We can also write by (4.5)

$$\lim_M \mathbb{E}\tilde{C}_{xy}/M = \Psi_{xy} \lim_M \mathbb{E}\tilde{C}_{yx}/M \text{ with } M \in \{K, N\}. \tag{4.20}$$

Now

$$\lim_{N\to\infty} \tilde{\Psi}_{xy} = \frac{\lim_N \mathbb{E}\tilde{C}_{xy}/N}{\lim_N \mathbb{E}\tilde{C}_{yx}/N} = \Psi_{xy}\frac{\lim_N \mathbb{E}\tilde{C}_{yx}/N}{\lim_N \mathbb{E}\tilde{C}_{yx}/N} = \Psi_{xy}.$$

□

4.5 Asymptotic Variances

We need to compute orders of $\pi^{st}_{xy|k}$ and $X^{st}_{xy|k}$ and to avoid confusion, we will omit the superscripts $s, t \in \{0, 1\}$ and write them instead as subscripts, e.g. π_{st} instead of $\pi^{st}_{xy|k}$. If used, then it will refer to the pair of items (x, y) and stratum k. Note $\pi^{st}_{xy|k} = \pi^{ts}_{yx|k}$. Also define $\tilde{L}_{xy} = \log \tilde{\Psi}_{xy}$.

As in the previous chapters, see Subsection 3.3.1 on page 95, we can write

$$\begin{aligned}
\lim_{M\to\infty} M \cdot \text{Var}^a(\tilde{L}_{xy}) &= \frac{1}{\Psi^2_{xy}} \lim_{M\to\infty} M \cdot \text{Var}^a(\tilde{\Psi}_{xy}) \\
&= \frac{1}{\Psi^2_{xy}} \frac{\lim_{M\to\infty} M \cdot \text{Var}^a(\tilde{\Omega}_{xy}/M)}{[\lim_{M\to\infty} \sum_{k=1}^K \mathbb{E}\tilde{c}_{yx|k}/M]^2} \\
&= \frac{\lim_{M\to\infty} \text{Var}^a(\tilde{\Omega}_{xy|k})/M}{[\lim_{M\to\infty} \sum_{k=1}^K \mathbb{E}\tilde{c}_{xy|k}/M]^2}
\end{aligned} \tag{4.21}$$

with $M \in \{K, N\}$. For the "sparse data" limiting model $\text{Var}^a(\tilde{\Omega}_{xy})/M = \lim_{K\to\infty} \frac{1}{K}$

$\sum_k \text{Var}(\tilde{\omega}_{xy|k})$ by Theorem 2.8.5 on page 75.

4.5.1 Computation of Var$(\tilde{\omega}_{xy|k})$

First we compute Var$(\tilde{\omega}_{xy|k})$.

$$
\begin{aligned}
&\text{Var}(\tilde{\omega}_{xy|k}) = \text{Var}(\tilde{c}_{xy} - \Psi\tilde{c}_{yx}) \\
&= \mathbb{E}(\tilde{c}_{xy} - \Psi\tilde{c}_{yx})^2 - [\mathbb{E}(\tilde{c}_{xy} - \Psi\tilde{c}_{yx})]^2 = \mathbb{E}(\tilde{c}_{xy} - \Psi\tilde{c}_{yx})^2 \\
&= \mathbb{E}\tilde{c}_{xy}^2 + \Psi^2\mathbb{E}\tilde{c}_{yx}^2 - 2\Psi\mathbb{E}\tilde{c}_{xy}\tilde{c}_{yx} \\
&= \frac{1}{n_k'^2}\{\mathbb{E}(X_{x|k}\bar{X}_{y|k} - X^{10}_{xy|k})^2 + \Psi^2\mathbb{E}(X_{y|k}\bar{X}_{x|k} - X^{10}_{yx|k})^2 \\
&- 2\Psi\mathbb{E}(X_{x|k}\bar{X}_{y|k} - X^{10}_{xy|k})(X_{y|k}\bar{X}_{x|k} - X^{10}_{yx|k})\} \\
&= \frac{1}{n_k'^2}\{(\mathbb{E}X_x^2\bar{X}_y^2 + \mathbb{E}X_{10}^2 - 2\mathbb{E}X_x\bar{X}_yX_{10}) + \Psi^2(\mathbb{E}X_y^2\bar{X}_x^2 + \mathbb{E}X_{01}^2 - 2\mathbb{E}X_y\bar{X}_xX_{01}) \\
&- 2\Psi(\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y - \mathbb{E}X_x\bar{X}_yX_{01} - \mathbb{E}X_y\bar{X}_xX_{10} + \mathbb{E}X_{10}X_{01})\} \\
&= \frac{1}{n_k'^2}\{\mathbb{E}X_x^2\bar{X}_y^2 + \mathbb{E}X_{10}^2 - 2\mathbb{E}X_x\bar{X}_yX_{10} + \Psi^2\mathbb{E}X_y^2\bar{X}_x^2 + \Psi^2\mathbb{E}X_{01}^2 - 2\Psi^2\mathbb{E}X_y\bar{X}_xX_{01} \\
&- 2\Psi\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y + 2\Psi\mathbb{E}X_x\bar{X}_yX_{01} + 2\Psi\mathbb{E}X_y\bar{X}_xX_{10} - 2\Psi\mathbb{E}X_{10}X_{01}\} \qquad (4.22)
\end{aligned}
$$

As noted previously, we assume $\mathbf{X} = (X_{11}, X_{10}, X_{01}, X_{00})$ follows a multinomial distribution with parameters n_k and $\boldsymbol{\pi} = (\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$ with $\pi_{11} + \pi_{10} + \pi_{01} + \pi_{00} = 1$. In order to compute Var$(\tilde{c}_{xy} - \Psi\tilde{c}_{yx})$, we need to consider the higher moments of the multinomial distribution. Assume a multinomial distribution with L possible outcomes and n independent trials; for each trial let p_i be the probability for outcome $i = 1, \ldots, L$ with $\sum_{i=1}^{L} p_i = 1$. Then let the random variables X_i be defined as the number of times outcome i was observed over the n trials. Define $\mathbf{X} = (X_1, X_2, \ldots, X_{k-1})$, $\mathbf{t} = (t_1, t_2, \ldots, t_{k-1})$, then $\mathbf{X}$ has the

following moment generation function

$$M_{\mathbf{X}}(\mathbf{t}) = (p_1 \exp(t_1) + p_2 \exp(t_2) + \cdots + p_{L-1} \exp(t_{L-1}) + p_L)^n.$$

We define $N_3 := nn'n''n'''$, $N_2 := nn'n''$, $N_1 := nn'$, $N_0 := n$ with $n' = n-1$, $n'' = n-2$ and $n''' = n-3$. Higher moments of the form $\mathbb{E}X_{i_1}^{s1} \cdot \cdots \cdot X_{i_m}^{s_m}$ $(m \leq K-1)$ are computed by

$$\frac{\partial^{(\sum_{i=1}^m s_i)} M_{\mathbf{X}}(\mathbf{t})}{\partial t_{i_1}^{s_1} \partial t_{i_2}^{s_2} \ldots \partial t_{i_m}^{s_m}}|_{\mathbf{t}=\mathbf{0}} = \mathbb{E}(X_{i_1}^{s_1} X_{i_2}^{s_2} \ldots X_{i_m}^{s_m}).$$

In this way, we yield the following moments up to the fourth order

$$\begin{aligned}
\mathbb{E}X_i &= N_0 p_i \\
\mathbb{E}X_i^2 &= N_1 p_i^2 + n p_i \\
\mathbb{E}X_i X_j &= N_1 p_i p_j \\
\mathbb{E}X_i^3 &= N_2 p_i^3 + 3N_1 p_i^2 + N_0 p_i \\
\mathbb{E}X_i^2 X_j &= N_2 p_i^2 p_j + N_1 p_i p_j \\
\mathbb{E}X_i X_j X_k &= N_2 p_i p_j p_k \\
\mathbb{E}X_i^4 &= N_3 p_i^4 + 6N_2 p_i^3 + 7N_1 p_i^2 + N_0 p_i \\
\mathbb{E}X_i^3 X_j &= N_3 p_i^3 p_j + 3N_2 p_i^2 p_j + N_1 p_i p_j \\
\mathbb{E}X_i^2 X_j^2 &= N_3 p_i^2 p_j^2 + N_2 (p_i^2 p_j + p_i p_j^2) + N_1 p_i p_j \\
\mathbb{E}X_i^2 X_j X_k &= N_3 p_i^2 p_j p_k + N_2 p_i p_j p_k \\
\mathbb{E}X_i X_j X_k X_l &= N_3 p_i p_j p_k p_l. \qquad (4.23)
\end{aligned}$$

For convenience, define $X_A := X_{10}$, $X_B := X_{01}$, $X_C := X_{11}$, $X_D := X_{00}$ to avoid confusion with the indices $s, t \in \{0, 1\}$, similarly for the π_{st}'s. Now we write n^2

and n as

$$n^2 = n''n''' + 5n'' + 4 = n'n'' + 3n' + 1 = nn' + n$$
$$n = n''' + 3 = n'' + 2 = n' + 1,$$

hence,

$$\begin{aligned} n^2N_1 = n^2n' &= N_3 + 5N_2 + 4N_1 & nN_2 = n^2n'n'' &= N_3 + 3N_2 \\ n^2N_0 = n^3 &= N_2 + 3N_1 + N_0 & nN_1 = n^2n' &= N_2 + 2N_1 \\ n^2 &= N_1 + N_0 & nN_0 = n^2 &= N_1 + N_0. \end{aligned} \tag{4.24}$$

Let$(\cdot)|_{N_i}$ denote the terms of $(\cdot)$ with factor N_i, for example $\mathbb{E}X_i^3X_j|_{N_3} = p_i^3p_j$. By applying (4.24) with (4.23), we derive the following higher moments (as shown in Appendix C on page 301)

$$\begin{aligned} \mathbb{E}X_A^2 &= N_1\pi_A^2 + N_0\pi_A \\ \mathbb{E}X_B^2 &= N_1\pi_B^2 + N_0\pi_B \\ \mathbb{E}X_AX_B &= N_1\pi_A\pi_B \\ \mathbb{E}X_x\bar{X}_yX_A &= N_2\pi_x\bar{\pi}_y\pi_A + N_1\{2\pi_A^2 + \pi_A - \pi_A\pi_B\} + N_0\pi_A \\ \mathbb{E}X_x\bar{X}_yX_B &= N_2\pi_x\bar{\pi}_y\pi_B + N_1\pi_A\pi_B \\ \mathbb{E}\bar{X}_xX_yX_B &= N_2\pi_y\bar{\pi}_x\pi_B + N_1\{2\pi_B^2 + \pi_B - \pi_B\pi_A\} + N_0\pi_B \\ \mathbb{E}\bar{X}_xX_yX_A &= N_2\pi_y\bar{\pi}_x\pi_A + N_1\pi_B\pi_A \\ \mathbb{E}X_x^2\bar{X}_y^2|_{N_3} &= \pi_x^2\bar{\pi}_y^2 \\ \mathbb{E}X_x^2\bar{X}_y^2|_{N_2} &= \pi_x\bar{\pi}_y(1 - \pi_B + 5\pi_A) \\ \mathbb{E}X_x^2\bar{X}_y^2|_{N_1} &= \pi_x\bar{\pi}_y + 4\pi_A^2 + 2\pi_A - 2\pi_A\pi_B \end{aligned}$$

$$
\begin{aligned}
\mathbb{E}X_x^2\bar{X}_y^2|_{N_0} &= \pi_A \\
\mathbb{E}X_y^2\bar{X}_x^2|_{N_3} &= \pi_y^2\bar{\pi}_x^2 \\
\mathbb{E}X_y^2\bar{X}_x^2|_{N_2} &= \pi_y\bar{\pi}_x(1-\pi_A+5\pi_B) \\
\mathbb{E}X_y^2\bar{X}_x^2|_{N_1} &= \pi_y\bar{\pi}_x+4\pi_B^2+2\pi_B-2\pi_A\pi_B \\
\mathbb{E}X_y^2\bar{X}_x^2|_{N_0} &= \pi_B \\
\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_3} &= \pi_x\pi_y\bar{\pi}_x\bar{\pi}_y \\
2\times\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_2} &= (\pi_x\bar{\pi}_y+\pi_y\bar{\pi}_x)(2\pi_A+2\pi_B+1)-2(\pi_A-\pi_B)^2-(\pi_A+\pi_B) \\
\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_1} &= \pi_x\bar{\pi}_y-\pi_A=\pi_y\bar{\pi}_x-\pi_B \\
\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_0} &= 0.
\end{aligned} \tag{4.25}
$$

Now we compute (4.22) by collecting all terms with factors $N_i, i = 0,1,2,3$ for ${n'_k}^2\text{Var}(\tilde{\omega}_{xy|k})$ separately.

$$
\begin{aligned}
&{n'}^2\text{Var}(\tilde{\omega}_{xy|k})|_{N_3} \\
&= \mathbb{E}X_x^2\bar{X}_y^2|_{N_3}+\mathbb{E}X_A^2|_{N_3}-2\mathbb{E}X_x\bar{X}_yX_A|_{N_3} \\
&+\Psi^2(\mathbb{E}X_y^2\bar{X}_x^2|_{N_3}+\mathbb{E}X_B^2|_{N_3}-2\mathbb{E}X_y\bar{X}_xX_B|_{N_3}) \\
&-2\Psi(\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_3}-\mathbb{E}X_x\bar{X}_yX_B|_{N_3}-\mathbb{E}X_y\bar{X}_xX_A|_{N_3}+\mathbb{E}X_AX_B|_{N_3}) \\
&= \pi_x^2\bar{\pi}_y^2+0+0+\Psi^2\pi_y^2\bar{\pi}_x^2+0+0-2\Psi\pi_x\pi_y\bar{\pi}_x\bar{\pi}_y+0+0+0 \\
&= \pi_x^2\bar{\pi}_y^2+\pi_x^2\bar{\pi}_y^2-2\pi_x^2\bar{\pi}_y^2 \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
&{n'}^2\text{Var}(\tilde{\omega}_{xy|k})|_{N_2} \\
&= \mathbb{E}X_x^2\bar{X}_y^2|_{N_2}+\mathbb{E}X_A^2|_{N_2}-2\mathbb{E}X_x\bar{X}_yX_A|_{N_2}
\end{aligned}
$$

$$
\begin{aligned}
&+ \Psi^2(\mathbb{E}X_y^2\bar{X}_x^2|_{N_2} + \mathbb{E}X_B^2|_{N_2} - 2\mathbb{E}X_y\bar{X}_xX_B|_{N_2}) \\
&- 2\Psi(\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_2} - \mathbb{E}X_x\bar{X}_yX_B|_{N_2} - \mathbb{E}X_y\bar{X}_xX_A|_{N_2} + \mathbb{E}X_AX_B|_{N_2}) \\
&= \pi_x\bar{\pi}_y(1 - \pi_B + 5\pi_A) + 0 - 2\pi_x\bar{\pi}_y\pi_A \\
&+ \Psi^2\pi_y\bar{\pi}_x(1 - \pi_A + 5\pi_B) + 0 - 2\Psi^2\pi_y\bar{\pi}_x\pi_B \\
&- \Psi(\pi_x\bar{\pi}_y + \pi_y\bar{\pi}_x)(2\pi_A + 2\pi_B + 1) + 2\Psi(\pi_A - \pi_B)^2 + \Psi(\pi_A + \pi_B) \\
&+ 2\Psi\pi_x\bar{\pi}_y\pi_B + 2\Psi\pi_y\bar{\pi}_x\pi_A - 0 \\
&= \pi_x\bar{\pi}_y\left[1 - \pi_B + 5\pi_A - 2\pi_A - (2\pi_A + 2\pi_B + 1) + 2\pi_A\right] \\
&+ \Psi^2\pi_y\bar{\pi}_x\left[1 - \pi_A + 5\pi_B - 2\pi_B - (2\pi_A + 2\pi_B + 1) + 2\pi_B\right] \\
&+ 2\Psi(\pi_A - \pi_B)^2 + \Psi(\pi_A + \pi_B) \\
&= 3\pi_x\bar{\pi}_y(\pi_A - \pi_B) + 3\Psi^2\pi_y\bar{\pi}_x(\pi_B - \pi_A) + 2\Psi(\pi_A - \pi_B)^2 + \Psi(\pi_A + \pi_B) \\
&= \Psi\{3\pi_y\bar{\pi}_x(\pi_A - \pi_B) + 3\pi_x\bar{\pi}_y(\pi_B - \pi_A) + 2(\pi_A - \pi_B)^2 + (\pi_A + \pi_B)\} \\
&= \Psi\{3(1 - \pi_x)\pi_y(\pi_A - \pi_B) - 3\pi_x(1 - \pi_y)(\pi_A - \pi_B) + 2(\pi_A - \pi_B)^2 + (\pi_A + \pi_B)\} \\
&= \Psi\{3(\pi_A - \pi_B)(\pi_y - \pi_x\pi_y - \pi_x + \pi_x\pi_y) + 2(\pi_A - \pi_B)^2 + (\pi_A + \pi_B)\} \\
&= \Psi\{3(\pi_A - \pi_B)(\pi_B - \pi_A) + 2(\pi_A - \pi_B)^2 + (\pi_A + \pi_B)\} \\
&= \Psi\{(\pi_A + \pi_B) - (\pi_A - \pi_B)^2\}
\end{aligned}
$$

$$
\begin{aligned}
&n'^2\mathrm{Var}(\tilde{\omega}_{xy|k})|_{N_1} \\
&= \mathbb{E}X_x^2\bar{X}_y^2|_{N_1} + \mathbb{E}X_A^2|_{N_1} - 2\mathbb{E}X_x\bar{X}_yX_A|_{N_1} \\
&+ \Psi^2(\mathbb{E}X_y^2\bar{X}_x^2|_{N_1} + \mathbb{E}X_B^2|_{N_1} - 2\mathbb{E}X_y\bar{X}_xX_B|_{N_1}) \\
&- 2\Psi(\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_1} - \mathbb{E}X_x\bar{X}_yX_B|_{N_1} - \mathbb{E}X_y\bar{X}_xX_A|_{N_1} + \mathbb{E}X_AX_B|_{N_1}) \\
&= \pi_x\bar{\pi}_y + 4\pi_A^2 + 2\pi_A(1 - \pi_B) + \pi_A^2 - 2(2\pi_A^2 + \pi_A - \pi_A\pi_B)
\end{aligned}
$$

$$
\begin{aligned}
&+ \Psi^2(\pi_y\bar{\pi}_x + 4\pi_B^2 + 2\pi_B(1-\pi_A) + \pi_B^2 - 2(2\pi_B^2 + \pi_B - \pi_B\pi_A)) \\
&- \Psi(\pi_x\bar{\pi}_y + \pi_y\bar{\pi}_x - \pi_A - \pi_B) + 2\Psi\pi_A\pi_B + 2\Psi\pi_B\pi_A - 2\Psi\pi_A\pi_B \\
&= \pi_x\bar{\pi}_y(1-1) + \Psi^2\pi_y\bar{\pi}_x(1-1) \\
&+ \pi_A^2(4+1-4) + \pi_A(2-2) + \pi_A\pi_B(-2+2) \\
&+ \Psi^2\{\pi_B^2(4+1-4) + \pi_B(2-2) + \pi_A\pi_B(-2+2)\} \\
&+ \Psi\{\pi_A + \pi_B + \pi_A\pi_B(2+2-2)\} \\
&= \pi_A^2 + \Psi^2\pi_B^2 + \Psi(\pi_A + \pi_B + 2\pi_A\pi_B)
\end{aligned}
$$

$$
\begin{aligned}
&{n'_k}^2 \mathrm{Var}(\tilde{\omega}_{xy|k})|_{N_0} \\
&= \mathbb{E}X_x^2\bar{X}_y^2|_{N_0} + \mathbb{E}X_A^2|_{N_0} - 2\mathbb{E}X_x\bar{X}_yX_A|_{N_0} \\
&+ \Psi^2(\mathbb{E}X_y^2\bar{X}_x^2|_{N_0} + \mathbb{E}X_B^2|_{N_0} - 2\mathbb{E}X_y\bar{X}_xX_B|_{N_0}) \\
&- 2\Psi(\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_0} - \mathbb{E}X_x\bar{X}_yX_B|_{N_0} - \mathbb{E}X_y\bar{X}_xX_A|_{N_0} + \mathbb{E}X_AX_B|_{N_0}) \\
&= \pi_A + \pi_A - 2\pi_A \\
&+ \Psi^2(\pi_B + \pi_B - 2\pi_B) \\
&- 2\Psi(0 - 0 - 0 + 0) \\
&= 0
\end{aligned}
$$

Overall we have

$$
\begin{aligned}
\mathrm{Var}(\tilde{\omega}_{xy|k}) = &\frac{N_2}{n'^2}\Psi\{(\pi_A + \pi_B) - (\pi_A - \pi_B)^2\} \\
&+ \frac{N_1}{n'^2}\{\pi_A^2 + \Psi^2\pi_B^2 + \Psi(\pi_A + \pi_B + 2\pi_A\pi_B)\}. \qquad (4.26)
\end{aligned}
$$

Note

$$n'^2\text{Var}(\tilde{\omega}_{xy|k})|_{N_2} \equiv \mathbb{E}X_x^2\bar{X}_y^2|_{N_2} + \Psi^2\mathbb{E}X_y^2\bar{X}_x^2|_{N_2} - 2\Psi\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_2}$$
$$\equiv n_k^2\text{Var}(\omega_{xy|k})|_{N_2}.$$

Greenland (1989) derived under J independent binomials

$$N_k{}^2\text{Var}(\omega_{xy|k}) = n_{x|k}n_{y|k}\{n'_{x|k}\pi_x\bar{\pi}_x + n'_{y|k}\pi_y\bar{\pi}_y\} + n_{x|k}n_{y|k}\{\pi_x\bar{\pi}_y + \pi_y\bar{\pi}_x\} \quad (4.27)$$

with $n_{x|k}$ referring to the totals of the xth binomial in the kth stratum and $N_k = \sum_x n_{x|k}$. For the sampling model of dependent binomials, the totals $n_{x|k}$ are all equal: $n_{1|k} = \cdots = n_{J|k} = n_k = N_k$.

Now we rewrite $\text{Var}(\tilde{\omega}_{xy|k})|_{N_2}$ as

$$\begin{aligned} n_k'^2\text{Var}(\tilde{\omega}_{xy|k})|_{N_2} &= (\pi_A + \pi_B) - (\pi_A - \pi_B)^2 \\ &= \pi_A + \pi_B - \pi_A^2 - \pi_B^2 + 2\pi_A\pi_B \\ &= \pi_A + \pi_B - \pi_A^2 - \pi_B^2 - 2\pi_C - 2\pi_B\pi_C - 2\pi_A\pi_C \\ &+ 2\pi_C + 2\pi_C + 2\pi_B\pi_C + 2\pi_A\pi_C - 2\pi_C + 2\pi_A\pi_B \\ &= (\pi_A + \pi_C)(1 - \pi_A - \pi_C) + (\pi_B + \pi_C)(1 - \pi_B - \pi_C) \\ &2(\pi_A + \pi_C)(\pi_B + \pi_C) - 2\pi_C \\ &= \pi_x\bar{\pi}_x + \pi_y\bar{\pi}_y + 2(\pi_x\pi_y - \pi_{xy}). \end{aligned} \quad (4.28)$$

Under independence $2(\pi_x\pi_y - \pi_{xy}) = 0$ and it becomes $\pi_x\bar{\pi}_x + \pi_y\bar{\pi}_y$, which is identical to $\text{Var}(\omega_{xy|k})|_{N_2}$ (neglecting factors) under the independent binomial model considered by Greenland (1989).

4.5.2 Large Stratum Limiting Variance

In Appendix D on page 309, we apply the delta method (Theorem 2.8.4) and show that

$$\lim_{N\to\infty}\frac{1}{N}\text{Var}^a(\tilde{\omega}_{xy|k}) = \alpha_k\{\pi_x\bar{\pi}_x + \pi_y\bar{\pi}_y + 2(\pi_x\pi_y - \pi_C)\}. \tag{4.29}$$

Computing $\lim_{N\to\infty}\frac{1}{N}\text{Var}(\tilde{\omega}_{xy|k})$ from (4.26) yields

$$\begin{aligned}
\frac{1}{N}\text{Var}(\tilde{\omega}_{xy|k}) &= \frac{N_2}{n'^2N}\Psi\{(\pi_A+\pi_B)-(\pi_A-\pi_B)^2\} \\
&\quad + \frac{N_1}{n'^2N}\{\pi_A^2+\Psi^2\pi_B^2+\Psi(\pi_A+\pi_B+2\pi_A\pi_B)\} \\
&= \frac{n''}{n'}\frac{n'}{n'}\frac{n}{N}\Psi\{(\pi_A+\pi_B)-(\pi_A-\pi_B)^2\} \\
&\quad + \frac{1}{n'}\frac{n'}{n'}\frac{n}{N}\{\pi_A^2+\Psi^2\pi_B^2+\Psi(\pi_A+\pi_B+2\pi_A\pi_B)\} \\
&\overset{N\to\infty}{\longrightarrow} \alpha_k\Psi\{(\pi_A+\pi_B)-(\pi_A-\pi_B)^2\}.
\end{aligned}$$

By (4.28), we see that $\lim_{N\to\infty}\frac{1}{N}\text{Var}(\tilde{\omega}_{xy|k}) \equiv \lim_{N\to\infty}\frac{1}{N}\text{Var}^a(\omega_{xy|k})$.
Equation (4.21) now becomes

$$\lim_{M\to\infty} M\cdot\text{Var}^a(\tilde{L}_{xy}) = \frac{\lim_{M\to\infty}\sum_k\frac{1}{M}\text{Var}(\tilde{\omega}_{xy|k})}{[\lim_{M\to\infty}\frac{1}{M}\sum_{k=1}^K\mathbb{E}\tilde{c}_{xy|k}]^2} \text{ for } M\in\{N,K\}. \tag{4.30}$$

This also proves that $(\sqrt{N}\cdot\omega_{xy|k}/N)^2$ is uniformly integrable by Theorem 2.8.7 on page 75. Hence by Lemma 1 on page 79, we could compute the asymptotic covariance directly by computing $\lim_{N\to\infty}\frac{1}{N}\text{Cov}(\tilde{\omega}_{xy|k},\tilde{\omega}_{wz|k})$, instead of applying the delta method. However, due to the complexity involved in the computation of the higher moments, we neither compute Cov($\tilde{\omega}_{xy|k}$, $\tilde{\omega}_{wz|k}$) nor propose any covariance estimator.

4.6 A Dually Consistent Variance Estimator of $\tilde{\Psi}$

Theorem 4.6.1.

$$\tilde{U}_{xyy} := \widehat{Var}(\log \tilde{\Psi}_{xy}) = \frac{\sum_k \frac{1}{n'^2}(X_A^2 - X_A)}{\tilde{C}_{xy}^2} + \frac{\sum_k \frac{1}{n'^2}(X_B^2 - X_B)}{\tilde{C}_{yx}^2} + \frac{\sum_k \frac{n''n'+2n'-1}{n'^2}(X_A + X_B) + \frac{2}{n'^2}X_A X_B - \frac{n''}{n'^2}(X_A - X_B)^2}{\tilde{C}_{xy}\tilde{C}_{yx}} \tag{4.31}$$

$\tilde{U}_{xyy}$ is dually consistent estimator of Var$(\log \tilde{\Psi}_{xy})$ and $\widehat{Var}(\tilde{\Psi}_{xy}) = \tilde{\Psi}_{xy}^2 \cdot \tilde{U}_{xyy}$ is a dually consistent estimators of Var$(\tilde{\Psi}_{xy})$.

Proof. The asymptotic variance (4.30) has denominator $\lim_{M\to\infty}(\lim_M \tilde{C}_{xy}/M)^2$. To show the dual consistency of $\tilde{U}_{xyy}$, we also reduce $\lim_{M\to\infty} M \cdot \tilde{U}_{xyy}$ such that the expression has the same denominator $(\lim_{M\to\infty}[\lim_M \tilde{C}_{xy}/M]^2)$ by applying formula (4.20). Then we only need to compare the numerators. By noting that $n''N_0 = n''n = N_1 - N_0$, the numerator of $\lim_{K\to\infty} K \cdot \tilde{U}_{xyy}$ is

$$\begin{aligned}
&= \lim_K \sum_k \frac{1}{n'^2 K}(X_A^2 - X_A) + \Psi^2 \sum_k \frac{1}{n'^2 K}(X_B^2 - X_B) \\
&+ \Psi \left\{ \lim_K \sum_k \frac{n''n' + 2n' - 1}{n'^2 K}(X_A + X_B) + \frac{2}{n'^2 K}X_A X_B - \frac{n''}{n'^2 K}(X_A - X_B)^2 \right\} \\
&= \lim_K \sum_k \frac{1}{n'^2 K}(\mathbb{E}X_A^2 - \mathbb{E}X_A + \Psi^2(\mathbb{E}X_B^2 - \mathbb{E}X_B)) + \Psi \frac{n''n' + 2n' - 1}{n'^2 K}(\mathbb{E}X_A + \mathbb{E}X_B) \\
&+ \Psi \lim_K \sum_k \frac{2}{n'^2 K}\mathbb{E}X_A X_B - \frac{n''}{n'^2 K}(\mathbb{E}X_A^2 + \mathbb{E}X_B^2 - 2\mathbb{E}X_A X_B) \\
&= \lim_K \sum_k \frac{1}{n'^2 K}\{N_1\pi_A^2 + N_1\Psi^2\pi_B^2\} + \Psi \frac{n''n' + 2n' - 1}{n'^2 K}N_0(\pi_A + \pi_B) \\
&+ \Psi \lim_K \sum_k \frac{2}{n'^2 K}N_1\pi_A\pi_B - \frac{n''}{n'^2 K}(N_1(\pi_A^2 + \pi_B^2 - 2\pi_A\pi_B) + N_0(\pi_A + \pi_B))
\end{aligned}$$

$$
\begin{aligned}
&= \lim_K \sum_k \Psi \frac{N_2}{n'^2 K}\{(\pi_A + \pi_B) - \pi_A^2 - \pi_B^2 + 2\pi_A\pi_B\} \\
&+ \lim_K \sum_k \frac{N_1}{n'^2 K}\{\pi_A^2 + \Psi^2\pi_B^2 + \Psi[2(\pi_A + \pi_B) + 2\pi_A\pi_B - (\pi_A + \pi_B)]\} \\
&+ \lim_K \sum_k \frac{N_0}{n'^2 K}(\pi_A + \pi_B)(-1 + 1) \\
&= \lim_K \sum_k \Psi \frac{N_2}{n'^2 K}\{(\pi_A + \pi_B) - (\pi_A - \pi_B)^2\} \\
&+ \lim_K \sum_k \frac{N_1}{n'^2 K}\{\pi_A^2 + \Psi^2\pi_B^2 + \Psi[\pi_A + \pi_B + 2\pi_A\pi_B]\},
\end{aligned}
$$

and the numerator of $\lim_{N\to\infty} N \cdot \tilde{U}_{xyy}$ is

$$
\begin{aligned}
&= \lim_N \sum_k \frac{1}{n^2 N}(X_A^2 - X_A) + \Psi^2 \lim_N \sum_k \frac{1}{n^2 N}(X_B^2 - X_B) \\
&+ \Psi \lim_N \sum_k \frac{n''n' + 2n' + 1}{n^2 N}(X_A + X_B) + \frac{2}{n^2 N}X_A X_B - \frac{n''}{n^2 N}(X_A - X_B)^2 \\
&= \lim_N \sum_k \frac{1}{N}(\frac{X_A^2}{n^2} + \Psi^2\frac{X_B^2}{n^2}) - \frac{1}{nN}(\frac{X_A}{n} + \Psi^2\frac{X_B}{n}) + \Psi\frac{n''n' + 2n' + 1}{n^2}\frac{n}{N}(\frac{X_A}{n} + \frac{X_B}{n}) \\
&+ \Psi \lim_N \sum_k \frac{2}{N}\frac{X_A}{n}\frac{X_B}{n} - \frac{n''}{N}(\frac{X_A^2}{n^2} + \frac{X_B^2}{n^2} - 2\frac{X_A}{n}\frac{X_B}{n}) \\
&= \sum_k 0 \cdot (\pi_A^2 + \Psi^2\pi_B^2) - 0 \cdot (\pi_A + \Psi^2\pi_B) + \Psi \cdot 1 \cdot \alpha_k(\pi_A + \pi_B) \\
&+ \Psi \sum_k 0 \cdot \pi_A\pi_B - \alpha_k(\pi_A^2 + \pi_B^2 - 2\pi_A\pi_B) \\
&= \sum_k \Psi\alpha_k\{\pi_A + \pi_B - (\pi_A^2 + \pi_B^2 - 2\pi_A\pi_B)\} \\
&= \sum_k \Psi\alpha_k\{\pi_A + \pi_B - (\pi_A - \pi_B)^2\}.
\end{aligned}
$$

By comparing these expressions with the numerator $\lim_{M\to\infty} \frac{1}{M}\sum_k \text{Var}(\omega_{xy|k})$ of (4.30) and (4.26), we conclude that $\tilde{U}_{xyy}$ is indeed dually consistent.

□

Remark 4.6.2. Greenland (1989) proposed the following variance estimator for $\log \Psi_{xy}$

$$U_{xyy}^{old} = \frac{\sum_{k=1}^{K} c_{xy|k} h_{xy|k}}{2C_{xy}^2} + \frac{\sum_{k=1}^{K} c_{xy|k} h_{yx|k} + c_{yx|k} d_{xy|k}}{2C_{xy}C_{yx}} + \frac{\sum_{k=1}^{K} c_{yx|k} h_{xy|k}}{C_{yx}^2} \qquad (4.32)$$

with $h_{xy|k} = (X_x + \bar{X}_y)/N_k$, which is dually consistent under independence of items (J independent binomials). U_{xyy}^{old} consists of 8 terms, estimating $N_k^2 \cdot \text{Var}^a(\omega_{xy|k})|_{N_2} = \pi_x\bar{\pi}_x + \pi_y\bar{\pi}_y = \pi_x\bar{\pi}_x(\pi_y + \bar{\pi}_y) + \pi_y\bar{\pi}_y(\pi_x + \bar{\pi}_x)$ by (4.27), which consists of 4 terms. Each of these 4 terms is estimated by averaging over two of the eight of U_{xyy}^{old}. Under dependence of items, $(n')^2 \cdot \text{Var}^a(\tilde{\omega}_{xy|k})|_{N_2} = \pi_x\bar{\pi}_x + \pi_y\bar{\pi}_y + 2(\pi_x\pi_y - \pi_{xy})$ by (4.28). Instead of constructing an estimator $\tilde{U}_{xyy}$ by matching (4.26) directly, we could also use $U_{xyy}^{old} + U_{xyy}^{add}$ instead, where U_{xyy}^{add} is an additional part to yield dual consistency. In such a way, we would incorporate Greenland's estimator in the new estimator, which has been accomplished in the previous chapter for a different sampling model (formula (3.20) on page 99). Such a construction by averaging over several terms to estimate one term would yield a more sufficient estimator and is favourable over $\tilde{U}_{xyy}$, given that the variances of the terms to be averaged over have about the same variance.

4.7 Example

Again we reconsider the UTI data in Table 1.1 on page 2, and again for simplicity, we exclude item E due to zero cell counts. The UTI data consists of 2 strata and 2 rows of multiple responses with 5 items. We simply merge row 1 (women without UTI history) with row 2 (with UTI history), to form stratified multiple response data with only one row of multiple responses and 5 items per stratum. The odds ratio estimators are defined for $2 \times J$ tables per stratum. The positive

responses of women with or without prior UTI history (given the age group) are now considered as the first row and the negative as the second row, forming such a 2×5 table.

The new MH approach gives $\{\tilde{L}_{AB}, \tilde{L}_{AC}, \tilde{L}_{AD}, \tilde{L}_{BC}, \tilde{L}_{BD}, \tilde{L}_{CD}\}$=$\{0.5045$, 1.315, 1.747,0.815, 1.240, $0.428\}$ with standard errors $\{0.14$,0.15 , 0.16 , 0.14 , 0.15 , $0.16\}$ by applying formula (4.31), whereas the old estimates not considering the dependence give $\{L_{AB}, L_{AC}, L_{AD}, L_{BC}, L_{BD}, L_{CD}\}$=$\{0.5050$, 1.323, 1.760,0.818, 1.245, $0.425\}$ with standard errors $\{0.18$,0.19 , 0.20 , 0.19 , 0.20 , $0.20\}$ by applying formula (4.32). Standard errors obtained by Greenland's formula (4.32) are higher than those obtained by the new formula (4.31).

We can only describe the relationship among contraceptives, e.g. which one is significantly more popular than the other. However, it might not be the main interest for the UTI data. We observe a significant difference for any two contraceptives. For instance, the odds for using contraceptive "oral" are $\exp(0.504) = 1.656$ times those for using contraceptive "lubricated condom". The formula and bootstrap (co)variance estimates with $B = 50,000$ can be found in Table 4.1.

4.8 Simulation Study

4.8.1 Simulation Scheme

As in the previous chapters, we conduct a simulation study to investigate the performance of the proposed new log odds ratio estimator $\tilde{L}_{xy}$ and its variance estimator $\tilde{U}_{xyy}$. Another aim is to double check the derived formulae, because of their complicated structure.

The simulation study compares $\log \tilde{\Psi}_{xy}$ with $\log \hat{\Psi}_{xy}$ and $\hat{\gamma}_{xy} := \hat{\beta}_x - \hat{\beta}_y$ from model (4.2), which we fit with GEE and an exchangeable correlation structure,

Table 4.1: The "bootstrap" with $B = 50,000$ (first line) and "formulae" (second line, first entry $\tilde{U}$, second entry U) (co)variance estimates of $\{L_{xy}, x, y = A, \ldots, D\}$, shown is $100\times$ (co)variance, * indicates that value is zero by definition, NA: no estimate available

	L_{AB}	L_{AC}	L_{AD}	L_{BC}	L_{BD}	L_{CD}
L_{AB}	5.85 1.94, 3.52	5.24 $NA, 0.81$	4.91 $NA, 0.73$	-0.65 $NA, -1.00$	-1.01 $NA, -0.90$	-0.38 $NA, 0.00^*$
L_{AC}		5.86 2.17, 3.72	5.06 $NA, 0.50$	0.57 $NA, 0.71$	-0.24 $NA, 0.00^*$	-0.84 $NA, -0.98$
L_{AD}			6.18 2.64, 4.10	0.11 $NA, 0.00^*$	1.20 $NA, 0.56$	1.06 $NA, 0.86$
L_{BC}				1.24 2.00, 3.57	0.78 $NA, 0.71$	-0.45 $NA, -1.13$
L_{BD}					2.22 2.38, 3.93	1.45 $NA, 0.99$
L_{CD}						1.91 2.65, 4.12

similar to Section 2.5 on page 59. The variance estimator for $\hat{\gamma}_{xy}$ is obtained by the formula $\text{Var}(\hat{\gamma}_{xy}) = \text{Var}(\hat{\beta}_x) + \text{Var}(\hat{\beta}_y) - 2\text{Cov}(\hat{\beta}_x, \hat{\beta}_y)$. As previously, we use the naive and the robust GEE variance, denoted by Var^{GEE}_{naive} and $\text{Var}^{GEE}_{robust}$ respectively. Furthermore, we include the bootstrap estimate of variance denoted by Var^{BT}. We only consider the case $J = 2$, because we only derived a variance estimator referring to two items. Without covariance estimators we are not able to compute a generalised (co-)variance estimator. Hence, considering more than 2 items is not sensible.

For given Ψ_{xy}, we fix the marginal probabilities of the first item to $\pi_{1|k} = 0.5$ and compute $\pi_{2|k}$ according to Ψ_{xy}: $\pi_{2|k} = 1/(1 + \Psi_{xy} \cdot \epsilon)$ with $\epsilon = \frac{1-\pi_{x|1}}{\pi_{x}|1} \equiv 1$ for $\pi_{1|k} = 0.5$. Again we use the odds ratio $\theta_{xy|ak}$

$$\theta_{xy|ak} = \frac{P(Y_x = 1, Y_y = 1|ak)P(Y_x = 0, Y_y = 0|ak)}{P(Y_x = 0, Y_y = 1|ak)P(Y_x = 1, Y_y = 0|ak)}$$

Table 4.2: Simulation results for L, $\tilde{L}$ and $\hat{\gamma}_{12}$

K,N_k,Ψ,θ-n_{MH},n_{GEE}	mean $L_{12}, \tilde{L}_{12}, \hat{\gamma}_{12}$	$^{100\cdot}$ mse $L_{12}, \tilde{L}_{12}, \hat{\gamma}_{12}$
5 , 20 , 1 , 4 - 0 , 657	$-0.0009, -0.0009, -0.0009$	$6.055, 5.854, 6.228$
5 , 20 , 4 , 1 - 0 , 0	$1.411, 1.411, 1.441$	$10.81, 10.81, 11.53$
5 , 20 , 4 , 10 - 0 , 0	$1.431, 1.402, 1.449$	$7.308, 6.959, 7.718$
20 , 5 , 1 , 4 - 0 , 2027	$-0.0001, -0.0001, -0.0001$	$6.662, 5.799, 7.67$
20 , 5 , 4 , 1 - 0 , 1861	$1.410, 1.411, 1.569$	$11.20, 11.42, 17.11$
20 , 5 , 4 , 10 - 0 , 3809	$1.528, 1.405, 1.633$	$10.36, 7.582, 15.87$
1 , 500 , 1 , 4 - 0 , 299	$-0.0004, -0.0004, -0.0004$	$1.116, 1.114, 1.116$
1 , 500 , 4 , 1 - 0 , 0	$1.389, 1.389, 1.389$	$2.000, 2.000, 2.000$
1 , 500 , 4 , 10 - 0 , 0	$1.390, 1.389, 1.390$	$1.352, 1.350, 1.352$
10 , 50 , 1 , 4 - 0 , 299	$-0.0008, -0.0007, -0.0008$	$1.107, 1.092, 1.120$
10 , 50 , 4 , 1 - 0 , 0	$1.388, 1.388, 1.402$	$2.063, 2.062, 2.128$
10 , 50 , 4 , 10 - 0 , 0	$1.398, 1.387, 1.406$	$1.322, 1.294, 1.367$
50 , 10 , 1 , 4 - 0 , 353	$-0.0001, -0.0001, -0.0001$	$1.180, 1.102, 1.264$
50 , 10 , 4 , 1 - 0 , 66	$1.390, 1.390, 1.466$	$2.116, 2.119, 2.979$
50 , 10 , 4 , 10 - 0 , 328	$1.450, 1.391, 1.497$	$1.826, 1.355, 2.764$
100 , 5 , 1 , 4 - 0 , 5804	$-0.0011, -0.0010, -0.0012$	$1.258, 1.095, 1.456$
100 , 5 , 4 , 1 - 0 , 6528	$1.392, 1.392, 1.554$	$2.215, 2.228, 5.558$
100 , 5 , 4 , 10 - 0 , 9174	$1.516, 1.392, 1.624$	$3.224, 1.370, 7.399$

log(4)= 1.386294

as a measure of dependence between items. For convenience, we assume a constant $\theta = \theta_{12|k}$ for all strata $k = 1, \ldots, K$. The stratum sample sizes $n_1 = \cdots = n_K$ are set constant. As before, the number of bootstrap samples is chosen as $B = 400$ and the number of simulations as $n = 10000$. We record the empirical variance (denoted by Var^{emp}) of the log odds estimator over $n = 10000$ simulations and consider it as the true variances. The number of simulations for which GEE did not converge is denoted by n_{GEE} and the number for which L (or $\tilde{L}$) could not be computed is denoted by n_{MH}. The simulation results are based only on those data sets for which both methods converged.

Table 4.3: Simulation results for the variance and covariance estimators

K , N_k , Ψ , θ n_{MH},n_{GEE}	100×mean $\text{Var}^{emp}(L_{12}), \text{Var}^{emp}(\tilde{L}_{12}), \text{Var}^{emp}(\hat{\gamma}_{12})$ $\text{Var}^{BT}(L_{12}), \text{Var}^{BT}(\tilde{L}_{12}), \text{Var}^{GEE}_{robust}(\hat{\gamma}_{12})$ $U_{122}, \tilde{U}_{122}, \text{Var}^{GEE}_{naive}(\hat{\gamma}_{12})$	100000×mse $\text{Var}^{BT}(L_{12}), \text{Var}^{BT}(\tilde{L}_{12}), \text{Var}^{GEE}_{robust}(\hat{\gamma}_{12})$ $U_{122}, \tilde{U}_{122}, \text{Var}^{GEE}_{naive}(\hat{\gamma}_{12})$
5 , 20 , 1 , 4	$6.056, 5.854, 6.228$	$-, -, -$
0 , 657	$5.569, 5.397, 5.744$	$10.48, 10.50, 9.574$
	$8.203, 5.497, 5.748$	$46.56, 9.266, 9.503$
5 , 20 , 4 , 1	$10.75, 10.75, 11.22$	$-, -, -$
0 , 0	$10.98, 11.00, 11.14$	$41.71, 44.69, 26.94$
	$10.63, 11.00, 11.14$	$14.40, 33.47, 26.28$
5 , 20 , 4 , 10	$7.108, 6.934, 7.322$	$-, -, -$
0 , 0	$7.376, 7.207, 7.265$	$32.62, 32.51, 22.73$
	$10.85, 6.975, 7.292$	$155.7, 23.12, 22.85$
20 , 5 , 1 , 4	$6.663, 5.8, 7.671$	$-, -, -$
0 , 2027	$5.71, 5.088, 7.126$	$17.42, 14.18, 14.56$
	$8.6, 5.839, 7.238$	$39.16, 14.17, 13.37$
20 , 5 , 4 , 1	$11.14, 11.36, 13.78$	$-, -, -$
0 , 1861	$10.36, 10.64, 13.35$	$67.54, 89.31, 47.89$
	$10.93, 12.31, 13.29$	$22.70, 90.67, 42.49$
20 , 5 , 4 , 10	$8.359, 7.549, 9.792$	$-, -, -$
0 , 3809	$8.217, 7.504, 8.338$	$52.14, 51.39, 374.1$
	$11.91, 7.560, 9.318$	$152.4, 38.40, 467.8$
1 , 500 , 1 , 4	$1.116, 1.114, 1.116$	$-, -, -$
0 , 299	$1.074, 1.073, 1.071$	$0.119, 0.119, 0.066$
	$1.603, 1.071, 1.071$	$2.379, 0.065, 0.066$
1 , 500 , 4 , 1	$2.000, 2.000, 2.000$	$-, -, -$
0 , 0	$2.074, 2.074, 2.061$	$0.435, 0.435, 0.196$
	$2.058, 2.064, 2.061$	$0.107, 0.200, 0.196$
1 , 500 , 4 , 10	$1.350, 1.349, 1.350$	$-, -, -$
0 , 0	$1.323, 1.322, 1.315$	$0.220, 0.220, 0.134$
	$2.059, 1.314, 1.315$	$5.09, 0.134, 0.134$
100 , 5 , 1 , 4	$1.258, 1.096, 1.456$	$-, -, -$
0 , 5804	$1.104, 0.9818, 1.412$	$0.347, 0.234, 0.138$
	$1.705, 1.153, 1.436$	$2.008, 0.140, 0.098$
100 , 5 , 4 , 1	$2.212, 2.226, 2.754$	$-, -, -$
0 , 6528	$1.872, 1.899, 2.602$	$1.59, 1.595, 0.515$
	$2.13, 2.360, 2.589$	$0.206, 0.704, 0.522$
100 , 5 , 4 , 10	$1.531, 1.368, 1.765$	$-, -, -$
0 , 9174	$1.487, 1.335, 1.763$	$0.363, 0.320, 0.249$
	$2.333, 1.450, 1.840$	$6.614, 0.320, 0.386$

4.8.2 Simulation Results

Table 4.2 shows the performance of the log odds ratio estimators. For the large stratum case ($K = 1$, $N_k = 500$), all three estimators L, $\tilde{L}$ and $\hat{\gamma}$ perform well. The sparser the data becomes, the worse $\hat{\gamma}$ is. Also, the higher the dependence, the better $\tilde{L}$. Only for independent items ($\theta = 1$), L and $\tilde{L}$ behave similarly well. Estimator $\tilde{L}$ stays almost unbiased for growing dependence, in contrast to L and $\hat{\gamma}$, which only seem unbiased under the large stratum case. Despite choosing the right correlation structure for GEE (there are only two items and one correlation parameter), we are surprised that GEE performs even worse than the ordinary MH estimator, which wrongly assumes independence between items. The bad performance cannot be explained by convergence problems, because the results are only shown for those simulations for which the MH and GEE methods converged.

Table 4.3 shows the performance of the variance estimators. As we assumed, $\tilde{U}_{xyy}$ is not a perfect estimator. Under dependence ($\theta \neq 1$), it performs better than U_{xyy}, which was to be expected. The bootstrap estimator $\text{Var}^{BT}(\tilde{L}_{12})$ performs better than $\tilde{U}_{xyy}$ only a few times: For low sample sizes (either $K = 20$, $N_k = 5$ or $K = 5$, $N_k = 20$) or when neither K nor N_k is large, otherwise $\tilde{U}_{xyy}$ is superior. This performance pattern is similar to the one observed in Section 3.7 on page 114 for the variance estimator $U_{xyy|ab}$ of $\log \hat{\Psi}_{xy|ab}$, where the bootstrap estimator of variance performed better than $U_{xyy|ab}$ for either $K = 20$, $N_k = 5$ or $K = 5$, $N_k = 20$.

The performance of $\text{Var}^{GEE}_{naive}(\hat{\gamma}_{12})$ and $\text{Var}^{GEE}_{naive}(\hat{\gamma}_{12})$ is quite similar, which can be explained by the fact that $J = 2$ items yield only 1 correlation parameter and therefore the working correlation is automatically correctly specified. For $J > 2$, we expect both the robust and naive variance to behave worse, however, the ro-

bust variance should outperform the naive variance estimator due to the increasing number of correlation parameters and the likely fact that the working correlation is wrong. In general, the performance of the GEE variance estimators is quite good, however, $\hat{\gamma}_{12}$ itself performed weakly, overestimating the actual log odds ratio resulting in a higher true (empirical) variance. Thus, a bad estimator with large variances cannot be recommended even though their variance estimators perform well in estimating this large variance.

The empirical variances of L_{xy} and $\tilde{L}_{xy}$ behave similar to the estimators L_{xy} and $\tilde{L}_{xy}$. Under independence, the empirical variance of L is smallest, whereas under dependence it is that of $\tilde{L}$.

We recommend the following: Under dependence, we clearly favour $\tilde{L}$ with $\tilde{U}_{xyy}$ as the new estimators over L and U_{xyy}, in contrast to independence, where we recommend L and U_{xyy} instead. Estimator $\hat{\gamma}$ only performs well under the large stratum case.

For more than $J = 2$ items, a new generalised variance estimator also could be constructed. However, for that we need new covariance estimators. This might be the subject of future research. Still, we can compute a generalised estimator of the odds ratio based on $\tilde{L}$, and the bootstrap method gives a fairly good variance estimator.

Future research might present a more efficient estimator $\tilde{U}_{xyy}$, as we outlined in Remark 4.6.2, as well as covariance estimators, that along with the asymptotic covariances are yet to be derived. We expect such a new estimator $\tilde{U}^{new}_{xyy}$ to perform similarly well as U_{xyy} under independence of items, but outclassing U_{xyy} under dependence.

Remark 4.8.1. There is, however, the question whether 2 items x and y are independent or not, which is essential in determining which estimator to be used. One

possibility is to construct a 2×2 contingency table, where the rows are the positive and negative responses for item x and where the columns are the positive and negative responses for item y. Testing independence is equivalent to testing whether the two-way interaction parameter of a saturated log-linear model is zero. Another possibility is to apply Pearson's chi-square test statistic which has one degree of freedom.

Chapter 5

Methods for Deletion Diagnostics for Homogenous Linear Predictor Models

5.1 Introduction

In a study published by Richert, Tokach, Goodband and Nelssen (1993), 262 farmers were questioned about their veterinary information sources. They were asked to tick one or more of the following items: (A) professional consultant, (B) veterinarian, (C) state or local extension service, (D) magazines, and (E) feed companies and reps. Agresti and Liu (2001) used "education" and "size" of farm as explanatory variables. Variable "education" has only two levels, whether the farmers had at least some college education or not, and "size" has the following levels: Less than 1,000, 1,000 to 2,000, 2,000 to 5,000, more than 5,000, which are the number of pigs they marketed annually. The example is referred to as farmers' data. The data can be cross-classified into a $2 \times 4 \times 5$ table (see Table 5.1) showing the total

number of positive responses for each item and for each education and farm size level.

Agresti and Liu (2001) considered several marginal modelling strategies, such as generalised estimation equations (GEE) (Liang and Zeger 1986), a generalisation of quasi-likelihood, and maximum likelihood (ML) estimation for generalised log-linear models (GLLM) (Lang and Agresti 1994), which both take the dependence between items into account. Preisser and Qaqish (1996) proposed regression diagnostics for GEE. They introduced simple explicit expressions for the effect (DBETA) and the influence (Cooks Distance) of deleting an arbitrary set of observations and some sub-cases, as the deletion of clusters and observations (responses) within a cluster. The Cook distance (Cook 1977) is a measure of influence for a set of observations to be deleted. Potential influential observations are high leverage points and outliers, but neither a high leverage point nor an outlier must be influential.

In this chapter, we want to investigate deletion diaginostics, such as the Cook distance, for HLP models (Lang 2005), an extension of GLLM, for analysing multiple response data, which has not been considered yet. The link function of a HLP model is many-to-one, in contrast to the one-to-one link function of the GEE method, making the deletion of observations different for both approaches. The deletion for HLP models becomes more complex and difficult. Our aim is to find a simplified but reliable method to calculate deletion diagnostics efficiently. We investigate three different but equivalent deletion methods for deleting a set of predictors. In particular, we propose a "delete=replace" method, which assigns dummy variables for the predictors/observations being deleted and another method which only deletes a set of predictors $\mathbf{z}_{ij}$ and the corresponding linear predictors η_{ij}. In most cases, these two methods are computationally sim-

pler than the method of direct deletion of joint observation and the corresponding vectors of predictors $\mathbf{z}_{ij}$. We do not only consider full solutions but also provide one-step approximations of DBETA and the Cook distance.

We proceed as follows. Section 5.2 introduces GEE and HLP models, Section 5.3 follows with introducing some existing GEE deletion diagnostic methods. For GEE we also investigate in which instances the "delete=replace" method and the method of direct deletion of observations yield identical model parameter estimates. Then we investigate deletion diagnostics for HLP models by considering the aforementioned 3 equivalent deletion methods. The methods are illustrated and compared by using the farmers' data (Section 5.4) and Section 5.5 finishes with discussing results and methods. We published these sections previously (Suesse and Liu 2008) in a similar but more compact form (8 pages only). Deletion diagnostics for generalised linear mixed models (Xiang et al. 2002), another possible modelling approach for multiple response data, are not considered here.

Table 5.1: Marginal table of farmers' veterinary information sources by education and number of pigs

Number of Positive Responses

		Information Source					
Education	Number of Pigs	A	B	C	D	E	Number of Subjects
No College	$< 1,000$	2	13	18	22	17	42
	$1,000 - 2,000$	2	15	10	11	15	27
	$2,000 - 5,000$	7	10	10	14	11	22
	> 5000	13	10	7	14	7	27
Some College	$< 1,000$	3	16	21	33	22	53
	$1,000 - 2,000$	2	10	15	22	10	42
	$2,000 - 5,000$	1	7	7	7	6	20
	> 5000	14	9	7	8	5	29
Total		44	90	95	131	93	262

5.2 Model Fitting

5.2.1 Marginal Models

We denote the J dimensional multiple response vector for subject i by $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})^T$, where y_{ij} represents the jth item response of subject $i = 1, \ldots, n$, which is 1 for a positive response and 0 for a negative response. The mean response $\mathbb{E}y_{ij} = \mu_{ij}$ equals the probability of a positive response π_{ij} for binary observations. We assume π_{ij} depends on the linear predictor $\eta_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\beta}_j$ through the link function $g_j(\cdot)$ by

$$g_j(\pi_{ij}) = g_j(\mu_{ij}) = \eta_{ij} = \mathbf{z}_{ij}^T \boldsymbol{\beta}_j. \tag{5.1}$$

Column vector $\mathbf{z}_{ij}$ is the ith subject contribution to the design matrix of the jth model depending on the ith subject's covariates, which are stored in column vector $\mathbf{x}_i$.

Let $\mathbf{Z}_i = \text{Diag}(\mathbf{z}_{i1}^T, \ldots, \mathbf{z}_{iJ}^T)$, also let $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iJ})^T$, similarly define $\boldsymbol{\mu}_i$, $\boldsymbol{\eta}_i$, g. We can also express (5.1) in vector form as

$$\mathrm{g}(\boldsymbol{\pi}_i) = \boldsymbol{\eta}_i = \mathbf{Z}_i \boldsymbol{\beta}, \tag{5.2}$$

with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_J^T)^T$ and where $\mathrm{g}(\boldsymbol{\pi}_i)$ stands for the column vector $(g_1(\pi_{i1}), \ldots, g_J(\pi_{iJ}))^T$. This modelling approach is called *marginal modelling* (Agresti and Liu 1999), because we model J univariate marginal distributions of $\mathbf{y}_i$.

We can write this in an even more compact form as

$$\mathrm{g}(\boldsymbol{\pi}) = \boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta}, \tag{5.3}$$

with $\mathbf{Z} = (\mathbf{Z}_1^T, \ldots, \mathbf{Z}_n^T)^T$, similarly $\boldsymbol{\pi}$ and $\boldsymbol{\eta}$. Here $\mathrm{g}(\boldsymbol{\pi})$ stands for the column

vector $(\mathbf{g}(\boldsymbol{\pi}_1)^T, \ldots, \mathbf{g}(\boldsymbol{\pi}_n^T))^T$.

Agresti and Liu (2001) discussed several models for the farmers' data. One of the best that fits well is

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \alpha_j + \beta_j \cdot s_i \tag{5.4}$$

with equally spaced scores $s_i = 1, 2, 3, 4$ depending on the ith subject size of farm ($< 1,000, \ldots, > 5,000$). For example, if farmer i marketed less than $1,000$ pigs a year, then $s_i = 1$. This model is linear in farm size and was called "LIN S".

The mean response model parameters $\boldsymbol{\beta}$ are of primary interest; in contrast, the association parameters or any other higher order parameters are only of very limited concern. One way of model fitting is to fit a generalised linear model (GLM) (McCullagh and Nelder 1989) for each of the J items separately, however, this maximum likelihood (ML) approach does not account for the dependence between items and yields less efficient parameter estimates. In the next two subsections, we consider the current two most common model fitting approaches. We introduce the model fitting approaches in detail in order to investigate deletion diagnostics based on these iterative algorithms in the sections thereafter.

5.2.2 Generalised Estimation Equations

In this subsection, we introduce the generalised estimating equations (GEE) approach developed by Liang and Zeger (1986). GEE is a multivariate extension of the quasi-likelihood approach (Wedderburn 1974). Let $\text{Var}(\mathbf{y}_i) = \mathbf{f}_i \cdot \phi^{-1}$ denote the variance of $\mathbf{y}_i$ with variance function $\mathbf{f}_i = f(\boldsymbol{\mu}_i)$ and scale or dispersion parameter ϕ. Let us assume that the univariate distributions of $\mathbf{y}_i$ are of the exponential family. Function $f(\cdot)$ gives the mean variance relationship which is

uniquely determined by the distribution within the class of the exponential family, for instance for binary observations $\text{Var}(y_{ij}) = f(\pi_{ij}) = \pi_{ij}(1-\pi_{ij})$. GEE estimates are obtained by computing the root of the GEE (or quasi-score equations)

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} (\mathbf{A}_i \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \tag{5.5}$$

where $\partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ is a $J_i \times p$ matrix, $\mathbf{A}_i = \sqrt{\mathbf{f}_i}$ is a $J_i \times J_i$ diagonal matrix with elements $\sqrt{\text{Var}(y_{ij})}$, $\mathbf{R}_i(\boldsymbol{\alpha})$ is the $J_i \times J_i$ correlation matrix for observation (cluster) i depending on parameter(s) $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)^T$, $J_i \leq J$ is the length of cluster i accounting for possibly different cluster lengths and define $J_+ := \sum_{i=1}^{n} J_i$. Here we use the general setting for GEE with varying cluster lengths J_i, for multiple response data we often have constant length $J_i = J$, e.g. $J_i = 5$ for the farmer's data. Let us define

$$\mathbf{W}_i = \mathbf{D}_i^{-1} \mathbf{A}_i^{-1} \mathbf{R}_i^{-1}(\boldsymbol{\alpha}) \mathbf{A}_i^{-1} \mathbf{D}_i^{-1},$$

with $\mathbf{D}_i = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\mu}_i$. Also, let $\mathbf{W} = \text{Diag}(\mathbf{W}_1, \ldots, \mathbf{W}_n)$, similarly defined for $\mathbf{D}$ and $\mathbf{R}$. Let $\mathbf{y} := (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$ denote all observations stacked in a single vector, denoted in a similar manner for all other defined vectors and matrices. If design matrix $\mathbf{Z}$ has full column rank, $\boldsymbol{\beta}$ can be estimated by iterated weighted least squares (Preisser and Qaqish 1996):

$$\hat{\boldsymbol{\beta}}^{new} = \left(\mathbf{Z}^T \mathbf{W} \mathbf{Z}\right)^{-1} \mathbf{Z}^T \mathbf{W} \mathbf{p} \tag{5.6}$$

with pseudo-observations $\mathbf{p} = \mathbf{Z}\hat{\boldsymbol{\beta}} + \mathbf{D}(\mathbf{y} - \boldsymbol{\mu})$, assuming the dispersion parameter ϕ and the correlation matrix $\mathbf{R}(\boldsymbol{\alpha}) = \text{Diag}(\mathbf{R}_1(\boldsymbol{\alpha}), \ldots, \mathbf{R}_n(\boldsymbol{\alpha}))$ are known and given. If unknown, they must be estimated consistently for every iterate. The

correlation matrix $\mathbf{R}(\hat{\boldsymbol{\alpha}})$ with an implicitly given correlation model is then called *working correlation*.

The GEE in (5.5) can also be expressed as

$$\sum_{i=1}^{n} \mathbf{U}_i = 0 \tag{5.7}$$

with $\mathbf{U}_i = \mathbf{M}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i = 0$, where $\mathbf{M}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}^T$, $\mathbf{V}_i = \mathbf{A}_i \mathbf{R}_i \mathbf{A}_i$ and $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$. If $\mathbf{R}$ is unknown, $\mathbf{V}_i$ is considered as the *working covariance*.

Theorem 5.2.1 (Liang and Zeger 1986 - "standard method"). *Under mild regularity conditions and given that :*

1. $\hat{\boldsymbol{\alpha}}$ *is* $n^{1/2}$ *consistent given* $\boldsymbol{\beta}$ *and* ϕ
2. $\hat{\phi}$ *is* $n^{1/2}$ *consistent given* $\boldsymbol{\beta}$,
3. $|\partial \hat{\boldsymbol{\alpha}} / \partial \phi|$ *is* $O_p(1)$

then $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ *is asymptotically multivariate Gaussian with zero mean and variance*

$$\lim_{n \to \infty} n \cdot \mathbf{J}_1^{-1} \mathbf{J}_2 \mathbf{J}_1^{-1}$$

where

$$\mathbf{J}_1 = \sum_{i=1}^{n} \mathbf{M}_i^T \mathbf{V}_i^{-1} \mathbf{M}_i \text{ and } \mathbf{J}_2 = \sum_{i=1}^{n} \mathbf{M}_i^T \mathbf{V}_i^{-1} Cov(\mathbf{y}_i) \mathbf{V}_i^{-1} \mathbf{M}_i.$$

The covariance $\text{Cov}(\mathbf{y}_i)$ is usually unknown and if replaced by $(\mathbf{y}_i - \boldsymbol{\mu})^T (\mathbf{y}_i - \boldsymbol{\mu})$ and substituting the parameters $\boldsymbol{\beta}$, ϕ and $\boldsymbol{\alpha}$ by their estimates, we yield the *robust* or *sandwich* variance estimate

$$\text{Cov}(\hat{\boldsymbol{\beta}})_{robust} = \hat{\mathbf{J}}_1^{-1} \hat{\mathbf{J}}_2 \hat{\mathbf{J}}_1^{-1}. \tag{5.8}$$

If the specified correlation $\mathbf{R}(\boldsymbol{\alpha})$ is correct, implying $\mathbf{J}_1 = \mathbf{J}_2$, then this robust variance simplifies to the *naive* variance:

$$\text{Cov}(\hat{\boldsymbol{\beta}})_{naive} = \phi^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i^T\mathbf{W}_i\mathbf{Z}_i\right)^{-1} = \phi^{-1}\left(\mathbf{Z}^T\mathbf{W}\mathbf{Z}\right)^{-1} \equiv \mathbf{J}_1^{-1}. \tag{5.9}$$

Estimation of $\boldsymbol{\alpha}$ and ϕ

Liang and Zeger (1986) suggested estimating the correlation and scale parameters from the Pearson residuals which are defined by

$$\hat{\rho}_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\widehat{\text{Var}}(y_{ij})}. \tag{5.10}$$

Then we can estimate ϕ by

$$\hat{\phi} = \sum_{i=1}^{n}\sum_{j=1}^{J}\hat{\rho}_{ij}^2/[J_+ - p]. \tag{5.11}$$

Given ϕ, the parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)^T$ are commonly estimated by the general approach

$$\hat{\alpha}_l = \phi^{-1}\sum_{i=1}^{n}\sum_{j_1,j_2\in S_l}\hat{\rho}_{ij_1}\hat{\rho}_{ij_2}/[N(n) - p] \tag{5.12}$$

where S_l is the set of indices j_1, j_2 for which the correlation parameters $R_{j_1j_2}(\boldsymbol{\alpha})$ of $\text{Corr}(\mathbf{y}_i) = \mathbf{R}(\boldsymbol{\alpha}) = (R_{j_1j_2})_{j_1,j_2=1}^{J}$ are assumed to be equal to the lth parameter α_l, in formula $S_l := \{j_1, j_2 : R_{j_1j_2} = \alpha_l\}$. The number $N(n)$ refers to the number of Pearson residuals the correlation is estimated over. However, the specific estimator depends on the choice of the correlation $\mathbf{R}(\boldsymbol{\alpha})$. We consider now some popular choices for the working correlation structure.

An *exchangeable* structure $R_{j_1 j_2} = \text{Corr}(y_{ij_1}, y_{ij_2}) = \alpha$ is estimated by

$$\hat{\alpha} = \phi^{-1} \sum_{i=1}^{n} \sum_{j_1 > j_2} \hat{\rho}_{ij_1} \hat{\rho}_{ij_2} / \left\{ \sum_{i=1}^{n} 1/2J_i(J_i - 1) - p \right\} \tag{5.13}$$

specifying $S = \{j_1 > j_2 : j_1, j_2 = 1, \ldots, J_i\}$ and $N(n) = \sum_{i=1}^{n} 1/2J_i(J_i - 1)$. We estimate the structure $\text{Corr}(y_{ij}, y_{i(j+1)}) = \alpha_j$ by

$$\hat{\alpha}_j = \phi^{-1} \sum_{i=1}^{n} \hat{\rho}_{ij} \hat{\rho}_{i(j+1)} / (n - p) \tag{5.14}$$

with the special case of *1-dependence* $\alpha = \alpha_j$ for $j = 1, \ldots, J - 1$, which can be estimated by

$$\hat{\alpha} = \sum_{j=1}^{J-1} \hat{\alpha}_j / (J - 1) \tag{5.15}$$

or to have the general form (5.12) by

$$\hat{\alpha} = \phi^{-1} \sum_{i=1}^{n} \sum_{j=1}^{J-1} \hat{\rho}_{ij} \hat{\rho}_{i(j+1)} / [n(J - 1) - p].$$

An *unstructured* correlation structure $R_{j_1 j_2} = \alpha_{j_1 j_2}$ is estimated by

$$\hat{\mathbf{R}} = \frac{\phi^{-1}}{n} \sum_{i=1}^{n} \mathbf{A}_i^{-1} \mathbf{r}_i \mathbf{r}_i^T \mathbf{A}_i^{-1} \tag{5.16}$$

or re-expressed as

$$\hat{\alpha}_{j_1 j_2} = \phi^{-1} \sum_{i=1}^{n} \hat{\rho}_{ij_1} \hat{\rho}_{ij_2} / n,$$

where the denominator can also be replaced by $n - p$ to match (5.12). Another popular and simple structure is the *independence* structure $R_{j_1 j_2} = 0$ for $j_1 \neq j_2$ and $R_{jj} = 1$. When this structure is chosen, items are treated as independent and the GEE are identical the likelihood equations when each of the J marginal

models is fitted as a GLM (McCullagh and Nelder 1989).

We note, $N = N(n)$ in (5.12) depends on the choice of the correlation $\mathbf{R}(\boldsymbol{\alpha})$, but also on the number of clusters n. For these structures, the estimation of ϕ is not required for the estimation of $\boldsymbol{\beta}$, because it cancels out in the calculation of $\mathbf{W}$. For more details, please refer to Liang and Zeger (1986) and its references therein.

Prentice (1988) and Zhao and Prentice (1990) considered the estimation of $\boldsymbol{\beta}$ and additionally the association parameters $\boldsymbol{\alpha}$, which are also modelled in terms of some explanatory variables, by extending the GEE approach. Our main focus is the estimation of $\boldsymbol{\beta}$, hence, the association parameters are regarded as nuisance parameters and the extended GEE approach is not further considered in this chapter. The choice and modelling of the correlation and the application of the extended GEE approach will be discussed in more detail for repeated multiple responses in Chapter 6.

Multivariate Generalised Linear Models

Let $\mathbf{x}_i$ be a column vector of covariates and let the observations $\mathbf{y}_i \in \mathbb{R}^J$ $(i = 1, \ldots, n)$ be conditionally independent and its distribution be from the simple exponential family. Then the ith contribution of the log-likelihood kernel $l = \sum_{i=1}^{n} l_i$ of a multivariate GLM (MGLM) can be written as follows (Fahrmeir and Tutz 2001, Chapter 2 and 3)

$$l_i = \{\mathbf{y}_i^T \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)\}/\phi, \tag{5.17}$$

where $\boldsymbol{\theta}_i$ is the natural parameter, $\mathbb{E}\mathbf{y}_i = \partial b(\boldsymbol{\theta_i})/\partial \boldsymbol{\theta_i} = \boldsymbol{\mu}_i$, $\text{Cov}(\mathbf{y}_i) = \boldsymbol{\Sigma}_i = \phi \partial^2 b(\boldsymbol{\theta_i})/\partial \boldsymbol{\theta_i} \partial \boldsymbol{\theta_i}^T$, ϕ is the scale or dispersion parameter. The first derivative can

be written as follows

$$\mathbf{U}_i = \frac{\partial l_i}{\partial \boldsymbol{\beta}} = \mathbf{M}_i^T \boldsymbol{\Sigma}_i^{-1} [\mathbf{y}_i - \boldsymbol{\mu}_i], \tag{5.18}$$

with $\mathbf{M}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}$.
The model is usually expressed as

$$\mathbf{g}(\boldsymbol{\mu}_i) = \mathbf{Z}_i \boldsymbol{\beta} \tag{5.19}$$

with vector valued link function $\mathbf{g}(\cdot)$ and design matrix $\mathbf{Z}_i$ depending on $\mathbf{x}_i$.
The likelihood equations

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{U}_i = \mathbf{0} \tag{5.20}$$

are solved to obtain ML estimates $\hat{\boldsymbol{\beta}}$. The expected information matrix $\mathcal{I}$ has the form

$$\mathcal{J} = \mathbb{E}\mathcal{I} = \sum_{i=1}^{n} \mathbf{M}_i \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i^T, \tag{5.21}$$

where $\mathcal{I} = \sum_{i=1}^{n} \partial^2 l_i / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T$ is the observed information matrix. Obviously, the likelihood equations (5.20) are identical to the GEE when $\boldsymbol{\Sigma}_i \equiv \mathbf{V}_i$ within the class of the simple exponential family, in other word, if the working correlation $\mathbf{R}_i$ (consequently also the working covariance) is correctly specified or is known, ML estimates and GEE estimates are identical. However, this does not apply for multiple response data, because the discrete underlying joint distribution is not fully specified by $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, and is not a member of the simple exponential family. A special sub-case of this equivalence between GEE and MGLM is the equivalence of a GLM and ordinary quasi-likelihood functions for univariate distributions within the simple exponential family.

5.2.3 Homogenous Linear Predictor Models

Table 5.1 shows the marginal counts of positive responses for each item $j = 1, \ldots, J$ and explanatory variables education and size. The observations from the underlying joint distribution can be found in Table 5.2. The first column shows each of the possible 2^J ($J = 5$) binary sequences j' of the form $(j'_1, \ldots, j'_J)$ with $j'_J \in \{0, 1\}$. We use j and j' to distinguish between marginal responses with index j referring to the items and joint observations with index j' referring to the 2^J outcomes. The other columns show $v_{kj'}$, the number of observations for sequence j' and for covariate setting $k = 1, \ldots, K$.

We can compute the marginal probabilities π_{kj} from the joint probabilities by a simple matrix multiplication $\pi_{kj} = \mathbf{b}_j^T \boldsymbol{\tau}_k$, in vector form $\boldsymbol{\pi}_k = \mathbf{B}\boldsymbol{\tau}_k$, where $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_c)^T$ is a matrix containing only zeros and ones. By using the same matrix $\mathbf{B}$, we can compute the marginal counts in Table 5.1 from the joint observations $v_{kj'}$ by $\mathbf{B}\mathbf{v}_k$. For instance, summing over the 16 last observations $v_{kj'}$ in Table 5.2 (observations for which response for item A was positive, i.e. $j'_1 = 1$) for setting $k = 1$ gives 2, the same number we find in Table 5.1 for setting $k = 1$ and item A. In this way, $\mathbf{b}_1$ is specified; the first 16 entries are zero and last 16 ones.

Note that the probability π_{kj} is identical to π_{ij} of model (5.1), if the ith observation has setting k. We can express model (5.2) in terms of joint probabilities as $\mathrm{g}(\boldsymbol{\pi}_k) = \mathrm{g}(\mathbf{B}\boldsymbol{\tau}_k) = \mathbf{Z}_k\boldsymbol{\beta}$ with $\mathrm{g} = (g_1, \ldots, g_J)^T$.

Assume a logistic link for all J marginal models, then (5.2) can be re-expressed as a generalised log-linear model (GLLM), which has the form $\mathbf{C} \log \mathbf{M}\, \mathbf{m}_k = \mathbf{Z}_k\boldsymbol{\beta}$, where $\mathbf{m}_k$ ($\mathbf{m}_k = v_{k+}\boldsymbol{\tau}_k$) contains the expected cell counts of the joint table (Table 5.2) and where $\mathbf{M}$ and $\mathbf{C}$ are some matrices.

The parameter estimates of the marginal model only specify the J mean responses π_{kj}, but they cannot uniquely determine the 2^J joint probabilities $\tau_{kj'}$,

Table 5.2: Joint table of farmers' veterinary information sources by education and number of pigs

Number of Joint Counts

j'	No College				Some College				
	Number of Pigs								
Binary Coding	$< 1,000$	1,000 -2,000	2,000 -5,000	$> 5,000$	$< 1,000$	1,000 -2,000	2,000 -5,000	$> 5,000$	Total
1=(00000)	0	0	0	0	0	0	0	0	0
2=(00001)	3	4	1	2	11	6	3	2	32
3=(00010)	7	4	4	6	14	14	4	4	57
4=(00011)	5	0	1	0	2	1	1	0	10
5=(00100)	7	3	1	0	6	7	4	2	30
6=(00101)	1	0	0	1	0	0	1	0	3
7=(00110)	4	0	0	0	4	2	0	1	11
8=(00111)	1	1	2	0	0	0	0	0	4
9=(01000)	5	5	2	1	2	4	5	4	28
10=(01001)	2	4	1	0	1	1	0	1	10
11=(01010)	0	0	0	1	1	0	0	0	2
12=(01011)	1	0	0	1	1	0	0	1	4
13=(01100)	0	0	0	1	0	1	0	0	2
14=(01101)	1	0	0	0	0	0	0	0	1
15=(01110)	1	0	0	0	4	2	0	0	7
16=(01111)	2	4	3	1	4	2	1	0	17
17=(10000)	1	0	3	6	0	1	0	10	21
18=(10001)	0	0	0	0	0	0	0	0	0
19=(10010)	0	0	0	1	0	0	0	0	1
20=(10011)	0	0	0	0	0	0	0	0	0
21=(10100)	0	0	0	0	0	0	0	1	1
22=(10101)	0	0	0	0	0	0	0	0	0
23=(10110)	0	0	0	1	0	1	0	0	2
24=(10111)	0	0	0	0	0	0	0	0	0
25=(11000)	0	0	0	2	0	0	0	0	2
26=(11001)	0	0	0	0	0	0	0	0	0
27=(11010)	0	0	0	0	0	0	0	0	0
28=(11011)	0	0	0	0	0	0	0	0	0
29=(11100)	0	0	0	0	0	0	0	1	1
30=(11101)	0	0	0	0	0	0	0	0	0
31=(11110)	0	0	1	1	0	0	1	1	4
32=(11111)	1	2	3	2	3	0	0	1	12
Number of Subjects v_{k+}	42	27	22	27	53	42	20	30	262

because this is a many-to-one relationship. Hence, maximising the likelihood kernel $\sum_{k=1}^{K} \mathbf{v}_k^T \log \mathbf{m}_k$ is not possible with standard ML procedures where the likelihood is expressed in terms of the model parameters.

An alternative method is maximising the likelihood subject to a system of constraints and Lagrange multipliers describing the underlying model. Lang and Agresti (1994) and Lang (1996) investigated ML estimation for GLLM using a variant of the constraint approach of Aitchison and Silvey (1958, 1960).

Lang (2004) developed a theory of the constraint approach for the broader class of multinomial-Poisson homogeneous (MPH) models. A sub-class of MPH models are homogeneous linear predictor models having the form $\mathbf{L}(\mathbf{m}_i) = \mathbf{Z}_i\boldsymbol{\beta}$, which were considered by Lang (2005). The class of linear predictor models considered by Bergsma (1997) is formally equivalent to HLP models.

According to Lang (2005), models being expressed in terms of $\boldsymbol{\tau}_k$ are automatically HLP models, hence, our marginal models having the form $\mathbf{L}(\mathbf{m}_k) = \mathbf{g}(\mathbf{B}\boldsymbol{\tau}_k) = \mathbf{Z}_k\boldsymbol{\beta}$ are within the class of HLP models. HLP models do not only allow the logistic link, but also any other smooth link functions $g_j(\cdot)$, such as the probit link, in contrast to GLLM. From $\boldsymbol{\tau}_k = \mathbf{m}_k/v_{+k}$ and $\pi_{kj} = \mathbf{b}_j^T\tau_k$, we can write model (5.4) as

$$\log\left(\frac{\mathbf{b}_j^T\mathbf{m}_k/v_{+k}}{1-(\mathbf{b}_j^T\mathbf{m}_i/v_{+k})}\right) = \alpha_j + \beta_j \cdot s_k$$

which is now of the form $\mathbf{L}(\mathbf{m}_k) = \mathbf{Z}_k\boldsymbol{\beta}$.

HLP models assume K independent samples (or strata) each from either a multinomial or Possion distribution. The *sampling plan* $(\mathbf{G}, \mathbf{G}_\mathbf{F}, \mathbf{v}_+)$ determines the distribution of cell counts $\mathbf{v} \in \mathbb{R}^d$, where the ith element of $\mathbf{v}$ contains the type i outcome, in our case $d = K \cdot 2^J$. The *population matrix* $\mathbf{G} \in \mathbb{R}^{d\times K}$ has elements $G_{ik} \in \{0, 1\}$ with conditions $G_{i+} = 1$ and $G_{+k} \geq 1$, in our case, $G_{+k} = 2^J$. If $G_{ik} = 1$ then the ith element of $\mathbf{v}$ (the type i outcome) from stratum k has a sample

size of v_{+k}, and if $G_{ik} = 0$ then stratum k does not contain the type i outcome. Matrix $\mathbf{G}_F$ is identical to matrix $\mathbf{G}$ if all strata are from a multinomial distribution, if however the kth column is omitted in $\mathbf{G}_F$, then the sample size v_{+k} is a Poisson variable. For multiple response data, matrix $\mathbf{G}$ equals $\mathbf{G}_F$ and is of size $K \cdot 2^J \times K$, where each column contains exactly 2^J ones and each row contains only one "one"; the remaining entries are zeros. Vector $\mathbf{v}_+ = (v_{1+}, \ldots, v_{K+})$ contains the fixed sample sizes for each of the K strata. The matrices $\mathbf{G}$ and $\mathbf{G_F}$ are needed later for the fitting algorithm. We leave further details to the interested reader (Lang 2005).

We express now a HLP model in the more compact form

$$\mathbf{L}(\mathbf{m}) = \mathbf{Z}\boldsymbol{\beta} \tag{5.22}$$

with $\mathbf{Z} = (\mathbf{Z}_1^T, \ldots, \mathbf{Z}_K^T)^T$ (similarly define $\mathbf{m}$) and $\mathbf{L}(\mathbf{m})$ standing for $(\mathbf{L}(\mathbf{m}_1)^T, \ldots, \mathbf{L}(\mathbf{m}_K)^T)^T$. Define $\boldsymbol{\xi} := \log \mathbf{m}$ by parameterising $\mathbf{m}$ to yield strictly positive estimates for $\mathbf{m}$. Let $\mathbf{U}$ be the orthogonal complement of $\mathbf{Z}$ (assuming $\mathbf{Z}$ has full column rank), then define $\mathbf{h}(\mathbf{m}) := \mathbf{U}^T\mathbf{L}(\mathbf{m})$ and $\mathbf{H} := \frac{\partial \mathbf{h}(\mathbf{m})^T}{\partial \mathbf{m}} = \frac{\partial \mathbf{L}^T}{\partial \mathbf{m}}\mathbf{U}$. From (5.22) follows $\mathbf{h}(\mathbf{m}) = \mathbf{0}$, the general form of multinomial-Poisson-homogeneous (MPH) models. The following iteration scheme was recommended by Lang (2005) based on the maximisation of the likelihood kernel $l(\boldsymbol{\xi}; \mathbf{v}) = \mathbf{v}^T\boldsymbol{\xi}$ subject to (5.22)

$$\hat{\boldsymbol{\theta}}^{new} = \hat{\boldsymbol{\theta}} - \mathbf{S}(\hat{\boldsymbol{\theta}})^{-1} s(\hat{\boldsymbol{\theta}}) \tag{5.23}$$

with

$$s(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{v} - \mathbf{e}^{\boldsymbol{\xi}} + \mathbf{H}(\boldsymbol{\xi})\boldsymbol{\lambda} \\ \mathbf{h}(\boldsymbol{\xi}) \end{bmatrix} \text{ and } \mathbf{S}(\boldsymbol{\theta}) = \begin{pmatrix} -\mathbf{D}(e^{\boldsymbol{\xi}}) & \mathbf{H}(\boldsymbol{\xi}) \\ \mathbf{H}(\boldsymbol{\xi})^T & \mathbf{0} \end{pmatrix},$$

where $\mathbf{D}(\mathbf{x}) := \text{Diag}(\mathbf{x})$, $\mathbf{H}(\boldsymbol{\xi}) = \partial \mathbf{h}(\boldsymbol{\xi})/\partial \boldsymbol{\xi}$ and $\boldsymbol{\theta} = (\boldsymbol{\xi}^T, \boldsymbol{\lambda}^T)^T$, and where $\boldsymbol{\lambda}$ de-

notes a vector of Lagrange multipliers. The inverse of the matrix $\mathbf{S}$ is expressed as (Lang 2005)

$$\mathbf{S}^{-1} = \begin{pmatrix} -\mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{D}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}^{-1} & \mathbf{D}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{D}^{-1}\mathbf{H})^{-1} \\ (\mathbf{H}^T\mathbf{D}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}^{-1} & (\mathbf{H}^T\mathbf{D}^{-1}\mathbf{H})^{-1} \end{pmatrix}. \quad (5.24)$$

Suppose a final unique solution $\hat{\mathbf{m}}$ exists, then it solves the restricted likelihood equations

$$\begin{bmatrix} \mathbf{v} - \mathbf{m} + \mathbf{D}(\mathbf{m})\mathbf{H}(\mathbf{m})\boldsymbol{\lambda} \\ \mathbf{h}(\mathbf{m}) \end{bmatrix} = \mathbf{0}$$

and the parameter estimates are computed by

$$\hat{\boldsymbol{\beta}} = \mathbf{R}_{\mathbf{Z}}\mathbf{L}(\hat{\mathbf{m}}) \quad (5.25)$$

with $\mathbf{R}_{\mathbf{Z}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$. The asymptotic covariance for $\hat{\boldsymbol{\beta}}$ is given by

$$\text{Cov}(\boldsymbol{\beta}) = \mathbf{R}_{\mathbf{Z}}\left(\hat{\mathbf{C}}_1 - \hat{\mathbf{C}}_1\mathbf{U}(\mathbf{U}^T\hat{\mathbf{C}}_1\mathbf{U})\mathbf{U}^T\hat{\mathbf{C}}_1 - \hat{\mathbf{C}}_2\right)\mathbf{R}_{\mathbf{Z}} \quad (5.26)$$

with

$$\mathbf{C}_1(\mathbf{m}) = \partial\mathbf{L}/\partial\mathbf{m}^T\mathbf{D}(\mathbf{m})\partial\mathbf{L}^T/\partial\mathbf{m}$$

and

$$\mathbf{C}_2(\mathbf{m}) = \partial\mathbf{L}/\partial\mathbf{m}^T\mathbf{D}(\mathbf{m})\mathbf{G}_F\mathbf{G}_F^T\mathbf{D}(\mathbf{m})\partial\mathbf{L}^T/\partial\mathbf{m}.$$

The asymptotic covariance for a HLP model of zero order with full column rank matrix $\mathbf{U}$ simplifies to

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{Z}^T\frac{\partial\mathbf{L}(\hat{\mathbf{m}})}{\partial\mathbf{m}^T}\mathbf{D}(\hat{\mathbf{m}})\frac{\partial\mathbf{L}(\hat{\mathbf{m}})^T}{\partial\mathbf{m}}\mathbf{Z}\right)^{-1}. \quad (5.27)$$

Marginal models of the form (5.2) depending on multinomial probabilities $\boldsymbol{\tau}_i$ through $\boldsymbol{\pi}_i$ are zero order HLP models and the orthogonal complement $\mathbf{U}$ can always be constructed to have full column rank. Consequently, formula (5.27) applies for these marginal models. The likelihood-ratio statistic is given by

$$G^2 = 2\mathbf{y}^T \log\left(\frac{\mathbf{y}}{\mathbf{m}}\right). \tag{5.28}$$

Note that ML estimates are not properly defined for zero cells in the joint table, as in our example, see Table 5.2. To overcome this problem, a tiny constant, e.g. 10^{-5}, is added to zero cell counts and the estimates are then called *extended ML estimates*.

5.3 Deletion Diagnostics for GEE and HLP models

In this section, we introduce some of the GEE diagnostics considered by Preisser and Qaqish (1996) and focus on a "deletion = replace" method. Then we concentrate on deletion diagnostics for HLP models also proposing the same "delete = replace" method.

5.3.1 GEE-Diagnostics

Let $\hat{\boldsymbol{\beta}}$ be the parameter estimate of all observations and $\hat{\boldsymbol{\beta}}_{[d]}$ be the estimate when a set d of observations is deleted, similarly for all other quantities, e.g. $\mathbf{y}_d$ denotes the set d of observations to be deleted, whereas $\mathbf{y}_{[d]}$ denotes all remaining observations not in set d. For given set d, the deletion diagnostics DBETA and Cook

distance (Cook 1977) are defined as

$$\text{DBETA}_{[d]} = \Delta_d \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[d]} \tag{5.29}$$

and

$$CD_{[d]} = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[d]})^T \text{Cov}(\hat{\boldsymbol{\beta}})^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[d]})/p. \tag{5.30}$$

Let matrices and vectors be partitioned in the following way

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{[d]} & \mathbf{W}_{[d]d} \\ \mathbf{W}_{d[d]} & \mathbf{W}_d \end{pmatrix}, \mathbf{y} = \begin{pmatrix} \mathbf{y}_{[d]} \\ \mathbf{y}_d \end{pmatrix}.$$

Now we list some results from Preisser and Qaqish (1996). GEE estimates are obtained by applying iterative algorithm (5.6). In the following, we denote the old parameter estimates by $\hat{\boldsymbol{\beta}}^{old}$ and the new (updated) estimates by $\hat{\boldsymbol{\beta}}^{new}$, which then become the old estimates in the next iteration. The final solution of the iteration scheme is denoted by $\hat{\boldsymbol{\beta}}^{final}$. The linear predictor is updated by $\hat{\boldsymbol{\eta}}^{new} = \mathbf{Z}\hat{\boldsymbol{\beta}}^{new} = \mathbf{Hp}$, where $\mathbf{H} = \mathbf{QW}$ and $\mathbf{Q} = \mathbf{Z}(\mathbf{Z}^T\mathbf{WZ})^{-1}\mathbf{Z}^T$. Thus, $\mathbf{H}$ can be seen as a projection matrix which maps the current iterate of the pseudo-observations $\mathbf{p}$ into the subspace of the linear predictor. The leverage of a cluster i can be defined as $tr(\mathbf{H}_i)$. The jth element on the diagonal of $tr(\mathbf{H}_i)$ is the leverage of the jth response in the ith cluster on the fitted value. Define the adjusted residuals by $\mathbf{e}_i = \mathbf{D}_i(\mathbf{y}_i - \boldsymbol{\mu}_i)$ and also let $\mathbf{V} = \mathbf{W}^{-1}$.

Theorem 5.3.1 (Preisser and Qaqish 1996)**.**

$$\hat{\boldsymbol{\beta}}_{[d]} \approx \hat{\boldsymbol{\beta}} - (\mathbf{Z}^T\mathbf{WZ})^{-1}\tilde{\mathbf{Z}}_d^T(\mathbf{W}_d^{-1} - \tilde{\mathbf{Q}}_d)^{-1}\tilde{\mathbf{e}}_d,$$

where

$$\tilde{\mathbf{Z}}_d := \mathbf{Z}_d - \mathbf{V}_{d[d]}\mathbf{V}_{[d]}^{-1}\mathbf{Z}_{[d]}, \qquad \tilde{\mathbf{Q}}_d := \tilde{\mathbf{Z}}_d(\mathbf{Z}^T\mathbf{W}\mathbf{Z})^{-1}\tilde{\mathbf{Z}}_d^T,$$
$$\tilde{\mathbf{e}}_d := \tilde{\mathbf{p}}_d - \tilde{\mathbf{Z}}_d\hat{\boldsymbol{\beta}} = \mathbf{e}_d - \mathbf{V}_{d[d]}\mathbf{V}_{[d]}^{-1}\mathbf{e}_{[d]}, \qquad \tilde{\mathbf{p}}_d = \mathbf{p}_d - \mathbf{V}_{d[d]}\mathbf{V}_{[d]}^{-1}\mathbf{p}_{[d]}.$$

Proposition 5.3.2 (Preisser and Qaqish 1996)**.** *The one-step approximation for* $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]}$ *is*

$$DBETAC_i := (\mathbf{Z}^T\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}_i^T(\mathbf{W}_i^{-1} - \mathbf{Q}_i)^{-1}\mathbf{e}_i$$

where i *refers to the* i*th cluster.*

For univariate observations DBETAC$_i$ equals

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]} \approx (\mathbf{Z}^T\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}_i^T\mathbf{W}_i^{1/2}(1-h_i)^{-1/2}r_{p_i} \tag{5.31}$$

with h_i being the ith diagonal element of $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{Z}(\mathbf{Z}^T\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{W}^{1/2}$ and $r_{p_i} = (y_i - \mu_i\{f_i(1-h_i)\})$, which is one type of Pearson residuals. The one-step approximation (5.31) was introduced by Pregibon (1981) for logistic regression and is also identical to the one Williams (1987) derived for GLM. Preisser and Qaqish (1996) also derived one-step approximations for $\Delta\hat{\boldsymbol{\beta}}$ deleting the jth response of the ith cluster. They also presented formulae for the Cook distance measuring the standardised influence on the linear predictor for deleting an arbitrary set of observations, for deleting the ith cluster and for deleting the jth response of the ith cluster. Again, as for the leverage, the one-step approximation (5.31) simplifies for univariate responses to the formula presented by Williams (1987) for GLM. Preisser and Qaqish (1996) also presented a one-step approximation for the studentised distance for the influence of the ith cluster on the overall fit.

Haslett and Haslett (2007) considered a "delete = replace" method, which replaces the deleted observations by its conditional best linear unbiased predictor (BLUP). However, their conditional residuals are of different nature than the marginal residuals $\mathbf{r}_i$. Now we consider another "delete = replace" method by augmenting the design matrix, which is equivalent to the deletion of a set d. Define the augmented design matrix $\tilde{\mathbf{Z}}$ by

$$\tilde{\mathbf{Z}} = \begin{pmatrix} \mathbf{Z}_{[d]} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d \end{pmatrix}. \tag{5.32}$$

The resulting parameter vector is of length $p + |d|$ and has the form

$$\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \tilde{\boldsymbol{\beta}}_{[d]} \\ \tilde{\boldsymbol{\beta}}_d \end{pmatrix}. \tag{5.33}$$

Design matrix $\tilde{\mathbf{Z}}$ assigns one parameter for each deleted observation, such that the added parameter vector $\tilde{\boldsymbol{\beta}}_d$ contains exactly $|d|$ parameters. Vector $\tilde{\boldsymbol{\beta}}_{[d]}$ is now independent of $\tilde{\boldsymbol{\beta}}_d$ and is only estimated over those observations that are not deleted, yielding parameter estimates as if the set d of observations is deleted. Additionally, the idea is that each parameter of $\tilde{\boldsymbol{\beta}}_d^{final}$ fits perfectly the assigned observations, such that $\mathbf{r}_d^{final}$ is zero. Let the method where observations are deleted be called *conventional* (deletion) method and the method which replaces design matrix $\mathbf{Z}$ by $\tilde{\mathbf{Z}}$ be referred to as "*delete = augment*" method. This method can also be seen as a "delete = replace" method, as mentioned by Haslett (1999, p.605), but we name it differently to distinguish between the two.

For the linear model, it is well known, that the conventional and the "delete = augment" method yield identical parameter estimates for generalised least squares

(GLS), that is $\tilde{\boldsymbol{\beta}}_{[d]}$ and $\hat{\boldsymbol{\beta}}_{[d]}$ are identical. For example Haslett (1999) mentioned that the here-called "delete = augment" is an alternative to his "delete = replace" approach. Also Peixoto and Lamotte (1989) noted that deleting "a case is equivalent to adding a dummy variable." Similarly, the parameters of vector $\tilde{\boldsymbol{\beta}}_d$ are also such added dummy variables. It is clear then, that the "delete = augment" method also works for GEE, because the iteration scheme (5.6) has the same form as for GLS. However, for GLS only one iteration of (5.6) is applied and it does not use pseudo-observations. It also does not need to estimate correlation structure parameters $\boldsymbol{\alpha}$ and the scale parameter ϕ. Therefore, we must carefully investigate in which instances the two methods, the conventional method and the "delete = augment" method, are identical or are at least approximately equal.

Before we formulate the theorem, which states in which instances the new iterates and final solutions of the two methods are equivalent, consider the following situations: (i) For some working correlations, e.g. for an exchangeable (5.13) or unstructured (5.16) correlation structure, the scale parameter cancels out in the computation of $\hat{\boldsymbol{\beta}}$ (with $\mathbf{W}$) and is redundant. (ii) Consider the deletion of whole clusters. Deletion of responses within a cluster are of different nature, because clusters are independent, but responses within a cluster are not.

Theorem 5.3.3. *1. Assume situation (ii) and that the old iterates of both methods are identical.*

(a) Suppose (i) is fulfilled and either the correlation structure $\mathbf{R}(\boldsymbol{\alpha})$ is known or estimation of $\tilde{\boldsymbol{\alpha}}$ is modified according to

$$\tilde{\alpha}_l^{modified} = \frac{[N(n) - p]}{[N(n - |d|) - p]} \cdot \tilde{\alpha}_l. \tag{5.34}$$

(b) If ϕ is unknown and (i) is not fulfilled, then we additionally modify the esti-

mation of $\tilde{\phi}$ according to

$$\tilde{\phi}^{modified} = \frac{J_+ - p}{J_+ - (\sum_{i \in d} J_i) - p} \cdot \tilde{\phi}. \tag{5.35}$$

Then the new iterates of both methods are identical such that

$$\tilde{\boldsymbol{\beta}}_{[d]}^{new} = \hat{\boldsymbol{\beta}}_{[d]}^{new}.$$

2. *Consider the same situations as under 1. but suppose starting values to be different. Then*

$$\tilde{\boldsymbol{\beta}}_{[d]}^{final} = \hat{\boldsymbol{\beta}}_{[d]}^{final}.$$

3. *Otherwise final solutions are only approximately equal*

$$\tilde{\boldsymbol{\beta}}_{[d]}^{final} \approx \hat{\boldsymbol{\beta}}_{[d]}^{final}.$$

Proof. We apply the formula for the inverse of a partitioned matrix, e.g. Searle (1982, p.261):

$$\begin{aligned}
&\left(\tilde{\mathbf{Z}}^T \mathbf{W} \tilde{\mathbf{Z}}\right)^{-1} \tilde{\mathbf{Z}}^T \mathbf{W} \mathbf{p} \\
&= \begin{pmatrix} \mathbf{Z}_{[d]}^T \mathbf{W}_{[d]} \mathbf{Z}_{[d]} & \mathbf{Z}_{[d]}^T \mathbf{W}_{[d]d} \\ \mathbf{W}_{d[d]} \mathbf{Z}_{[d]} & \mathbf{W}_d \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Z}_{[d]}^T \mathbf{W}_{[d]} \mathbf{p}_{[d]} + \mathbf{Z}_{[d]}^T \mathbf{W}_{[d]d} \mathbf{p}_d \\ \mathbf{W}_{d[d]} \mathbf{p}_{[d]} + \mathbf{W}_d \mathbf{p}_d \end{pmatrix} \\
&= \begin{pmatrix} (\mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{Z}_{[d]})^{-1} & (\mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{Z}_{[d]})^{-1} \mathbf{Z}_{[d]}^T \mathbf{W}_{[d]d} \mathbf{W}_d^{-1} \\ \mathbf{W}_d^{-1} \mathbf{W}_{d[d]} \mathbf{Z}_{[d]} \mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{Z}_{[d]})^{-1} & \mathbf{W}_d^{-1} \mathbf{W}_{d[d]} \mathbf{Z}_{[d]} (\mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{Z}_{[d]})^{-1} \mathbf{Z}_{[d]}^T \mathbf{W}_{[d]d} \mathbf{W}_d^{-1} \end{pmatrix} \\
&\quad \begin{pmatrix} \mathbf{Z}_{[d]}^T \mathbf{W}_{[d]} \mathbf{p}_{[d]} + \mathbf{Z}_{[d]}^T \mathbf{W}_{[d]d} \mathbf{p}_d \\ \mathbf{W}_{d[d]} \mathbf{p}_{[d]} + \mathbf{W}_d \mathbf{p}_d \end{pmatrix}
\end{aligned}$$

$$= \begin{pmatrix} (\mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{Z}_{[d]})^{-1} \mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{p}_{[d]} \\ -\mathbf{W}_d^{-1} \mathbf{W}_{d[d]} \left(\mathbf{Z}_{[d]} (\mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{Z}_{[d]})^{-1} \mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{p}_{[d]} - \mathbf{p}_{[d]} \right) + \mathbf{p}_d \end{pmatrix}.$$

We note that $\mathbf{p} = \tilde{\mathbf{Z}} \tilde{\boldsymbol{\beta}}^{old} + \mathbf{D}\mathbf{r}$ can be decomposed into $\mathbf{p}_{[d]} = \mathbf{Z}_{[d]} \tilde{\boldsymbol{\beta}}_{[d]}^{old} + \mathbf{D}_{[d]} \mathbf{r}_{[d]}$ and $\mathbf{p}_d = \tilde{\boldsymbol{\beta}}_d^{old} + \mathbf{D}_d \mathbf{r}_d$, because $\mathbf{D}$ is a $J_+ \times J_+$ diagonal matrix. So, we derive

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{[d]}^{new} &= (\mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{Z}_{[d]})^{-1} \mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{p}_{[d]} \\ &= (\mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{Z}_{[d]})^{-1} \mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \tilde{\boldsymbol{\beta}}_{[d]}^{old} + \mathbf{D}_{[d]} \mathbf{r}_{[d]} \end{aligned} \tag{5.36}$$

and

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_d^{new} &= -\mathbf{W}_d^{-1} \mathbf{W}_{d[d]} \left(\mathbf{Z}_{[d]} (\mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{Z}_{[d]})^{-1} \mathbf{Z}_{[d]}^T \mathbf{V}_{[d]}^{-1} \mathbf{p}_{[d]} - \mathbf{p}_{[d]} \right) + \mathbf{p}_d \\ &= -\mathbf{W}_d^{-1} \mathbf{W}_{d[d]} \left(\mathbf{Z}_{[d]} \tilde{\boldsymbol{\beta}}_{[d]}^{old} - [\mathbf{Z}_{[d]} \tilde{\boldsymbol{\beta}}_{[d]}^{old} + \mathbf{D}_{[d]} \mathbf{r}_{[d]}] \right) + \tilde{\boldsymbol{\beta}}_d^{old} + \mathbf{D}_d \mathbf{r}_d \\ &= \mathbf{W}_d^{-1} \mathbf{W}_{d[d]} \mathbf{D}_{[d]} \mathbf{r}_{[d]} + \tilde{\boldsymbol{\beta}}_d^{old} + \mathbf{D}_d \mathbf{r}_d. \end{aligned} \tag{5.37}$$

Generally, we assume that there exists only a unique set of solutions for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ and that independently of the starting values the algorithm will converge to this unique set of solutions. Otherwise the algorithm would provide different solutions for different starting values and considering in which instances solutions are identical would be meaningless. First, let us assume a set d of clusters is deleted. It follows $\mathbf{W}_{d[d]} = \mathbf{W}_{[d]d} = \mathbf{0}$. Consider the case (1a) and that the correlation is known, which determines in each step a unique $\mathbf{V}$ only depending on the current mean $\boldsymbol{\mu}$. From (5.6) and (5.36) follows $\tilde{\boldsymbol{\beta}}_{[d]}^{new} \equiv \hat{\boldsymbol{\beta}}_{[d]}^{new}$. Now let us assume that the correlation is unknown and must be estimated according to (5.12). Unless we start with starting value $\tilde{\boldsymbol{\beta}}_d^{old}$, such that $\mathbf{r}_d^{old} = \mathbf{0}$, the residuals from set d will contribute to the estimation of the correlation parameters. However, the

parameters $\tilde{\boldsymbol{\beta}}_d$ will be updated to obtain finally $\mathbf{r}_d^{final} = \mathbf{0}$. When this is achieved, only residuals from the set of clusters that are not deleted will contribute to the estimation of the correlation parameters. Now it is clear that $N(n)$ is too large in (5.12) and needs to be changed according to (5.34). Then it is obvious from (5.36), that both methods will converge to the same parameter estimates.

(1b): The same modification for the estimation of ϕ is obviously needed if ϕ is unknown and condition (i) is not fulfilled.

(2) If starting values are not identical, then final solutions for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ are identical, because we assumed that the algorithm converges to a set of unique solutions and the iteration schemes for both methods use exactly the same formulae for $\tilde{\boldsymbol{\beta}}_{[d]}^{new}$ and $\hat{\boldsymbol{\beta}}_{[d]}^{new}$.

(3): We consider the case of deletion of single components of the clusters. In contrast, to cluster deletion, we have now $\mathbf{W}_{d[d]} = \mathbf{W}_{[d]d} \neq \mathbf{0}$. Hence, we see from (5.37), that there is an additional term $\mathbf{W}_d^{-1}\mathbf{W}_{d[d]}\mathbf{D}_{[d]}\mathbf{r}_{[d]}$ contributing to $\tilde{\boldsymbol{\beta}}_d^{new}$. Thus, generally $\mathbf{r}_d^{final} \neq \mathbf{0}$, which implies that $\mathbf{r}_d^{final}$ contributes to the estimation of the correlation parameters. These correlation parameters will be slightly different at final convergence for the two methods. Consequently $\mathbf{V}$ will also differ for both methods. Thus, the two methods will provide different solutions, however, the difference between correlation parameters and the scale parameters for both methods is small, but the basic formula to obtain estimates $\boldsymbol{\beta}$ remain the same yielding only slightly different final solutions $\tilde{\boldsymbol{\beta}}_{[d]}^{final}$ and $\hat{\boldsymbol{\beta}}_{[d]}^{final}$. The larger n is, the smaller is the difference, because GEE yields consistent estimates even if the working correlation structure is wrongly specified. Here, the working correlation is the equal for both methods, only the estimates of the correlation parameters are slightly different. □

Remark 5.3.4. In practical terms, under situation 3. of Theorem 5.3.3, when the

two methods only provide approximately equal final solutions, the difference between $\tilde{\boldsymbol{\beta}}_{[d]}^{final}$ and $\hat{\boldsymbol{\beta}}_{[d]}^{final}$ is relatively small. For example, if the correlation parameter for an exchangeable structure is not updated according to (5.34) for model "LIN S", then the Euclidean norm of the difference between the final parameter estimates of the two methods gives $\approx 5 \cdot 10^{-3}$. The difference between one-step approximations is far bigger, for the same model "LIN S" yields differences of around 0.5.

Remark 5.3.5. ("Sparse Data") How do these two methods perform under a sparse data situation? GEE yields consistent estimates even if the (working) correlation is wrongly specified. Under situation 3. of Theorem 5.3.3, the estimates of the correlation parameters will be slightly different, however the working correlation is still the same. Under very sparse data, the impact of slightly different correlation estimates will be higher than for large n. Therefore we would expect that the "delete=augment" method might yield more inaccurate results under this situation. Future research might clarify how reliable the "delete=augment" method is under such a sparse data situation.

5.3.2 HLP Diagnostics

First, we point out some differences between the GEE and HLP approaches. Marginal model (5.2) refers to n observations $\mathbf{y}_i$ of length J (assuming $J_i = J$). The total length of $\mathbf{y}$ and the corresponding mean vector $\boldsymbol{\pi}$ is $n \cdot J$. The vector of link functions $\mathbf{g}(\boldsymbol{\pi})$ is a one-to-one mapping from $\mathbb{R}^{n \cdot J}$ to $\mathbb{R}^{n \cdot J}$. To apply the HLP model methodology, we must express the marginal model in terms of expected joint table frequencies $\mathbf{m}_k$. For each setting $k = 1, \ldots, K$, there are 2^J such frequencies, so that the overall model (5.22) refers to the vector $\mathbf{m}$ of length $K \times 2^J$. The function $\mathbf{L}(\mathbf{m})$ is not one-to-one but many-to-one and maps from $K \times 2^J$ to $K \times J$.

Both approaches still refer to the same model, if the ith observation lies in the kth group (group k comprises of all observations with covariate setting k), then π_{ij} and π_{kj} are identical. The vector $\boldsymbol{\pi}$ in model (5.3) has v_{k+} entries for π_{kj}. In contrast, for the HLP approach, the model function $\mathbf{L}(\mathbf{m}_k)$, which can also be expressed as $\mathbf{g}(\mathbf{B}\boldsymbol{\tau}_k)$, refers only to one such $\pi_{kj} = \mathbf{b}_j^T \boldsymbol{\tau}_k$.

For the farmers' data ($J = 5$, $n = 262$, $K = 8$), $\mathbf{g}$ in (5.3) maps from 1310 to $1310(= 5 \times 262)$, but $\mathbf{L}$ maps from $256(= 2^5 \times 8)$ to $40(= 8 \times 5)$. The set d of the GEE diagnostics refers to any of the 1310 responses.

For the HLP approach, we must first consider what is to be deleted. The HLP model function $\mathbf{L}$ links the linear predictor $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta}$ with the expected cell counts $\mathbf{m}$ of the joint table. We distinguish between the dimension of the argument $\mathbf{m}$ of $\mathbf{L}(\mathbf{m})$ and the dimension of the linear predictor. Let index d refer now to any of the $K \times J$ components of $\boldsymbol{\eta} = \mathbf{Z}\boldsymbol{\beta}$, which can be considered as a marginal index set, because the linear predictor $\boldsymbol{\eta}$ predicts the marginal probabilities $\boldsymbol{\pi}$ through one-to-one link function $\mathbf{g}$. Let index d' refer to any of the joint observations $\mathbf{v}$ of length $K \times 2^J$, which is considered as a joint index set. When we say "delete a set d of predictors", we mean that we delete the corresponding rows of predictors of design matrix $\mathbf{Z}$ (or equivalently the components of $\boldsymbol{\eta}$) and the components of link function $\mathbf{L}(\cdot)$.

The deletion of the set d' might be equivalent to the deletion of the set d. For example, deleting the joint observations with farm size $< 1,000$ and some college education (setting $k = 5$) is equivalent to deleting the corresponding $1 \times J = 5$ of the total $K \times J = 40$ predictors. However, for the same model, the deletion of an item which can be accomplished by deleting a set d of $1 \times K = 8$ predictors, cannot be achieved by deleting a set d' of joint observations. Instead the joint observations must be manipulated so that the data has one item removed. In

general, the deletion of a set d might be more meaningful than deleting of a set d', because it shows the influence of a set of predictors.

Assume a set d of linear predictors is deleted, along with the corresponding set of marginal observations which are determined by $\mathbf{B}\mathbf{v}_k$. Then we consider three possible deletion methods. The first is the conventional method, which is deleting or manipulating the joint observations, whichever applies, so that the corresponding set d of predictors are deleted (method 1). The second method (method 2) is the "replace = augment" method, which leaves the observations untouched and only replaces $\mathbf{Z}$ by $\tilde{\mathbf{Z}}$ defined in (5.32). We point out for GEE we had $\mathbf{Z} = (\mathbf{Z}_1^T, \ldots, \mathbf{Z}_n^T)^T$, but for HLP we defined $\mathbf{Z}$ as $\mathbf{Z} = (\mathbf{Z}_1^T, \ldots, \mathbf{Z}_K^T)^T$. The third method (method 3) only deletes the set d of predictors $\mathbf{z}_{ij}$ and the corresponding components of $\mathbf{L}(\mathbf{m})$, but leaves the joint observations untouched.

Consider deleting only one item. Method 1 reduces the number of items of the joint observations to 4. The function $\mathbf{L}$ and matrix $\mathbf{Z}$ must also be changed, such that $\mathbf{L}$ now maps from $K \cdot 2^{J-1}$ to $K \cdot (J-1)$. Method 2 only modifies the design matrix $\mathbf{Z}$ and assigns dummy variables. The third method does not change the joint observations and also does not change the function $\mathbf{L}$ itself; it only deletes the components (or rows) of $\mathbf{L}$ and $\mathbf{Z}$ referring to the item to be deleted, such that $\mathbf{L}$ maps now from $K \cdot 2^J$ to $K \cdot (J-1)$.

As previously, we use the hat symbol for estimates/quantities of the first method (e.g. $\hat{\beta}$), the tilde symbol for the second (e.g. $\tilde{\beta}$) and for the third, we use the bar symbol (e.g. $\bar{\boldsymbol{\beta}}$).

Theorem 5.3.6. *Assume a unique solution always exists. (a) Let us assume deleting set d is equivalent to deleting a set d'. Then the three deletion methods are equivalent in the sense that*

$$\hat{\boldsymbol{\beta}}_{[d]}^{new} = \tilde{\boldsymbol{\beta}}_{[d]}^{new} = \bar{\boldsymbol{\beta}}_{[d]}^{new}.$$

(b) If the assumption of equivalence of deletion of sets d and d' does not hold, the "replace = augment" and the third method are still equivalent

$$\tilde{\boldsymbol{\beta}}_{[d]}^{new} = \bar{\boldsymbol{\beta}}_{[d]}^{new},$$

and all three method yield equal final solutions

$$\hat{\boldsymbol{\beta}}_{[d]}^{final} = \tilde{\boldsymbol{\beta}}_{[d]}^{final} = \bar{\boldsymbol{\beta}}_{[d]}^{final}.$$

Proof. Equivalence of methods 1 and 2: First, we show that the iterative scheme (5.23) produces equivalent next iterates $\tilde{\boldsymbol{\xi}}_{[d']} = \hat{\boldsymbol{\xi}}_{[d']}$ with

$$\tilde{\boldsymbol{\xi}} = \begin{pmatrix} \tilde{\boldsymbol{\xi}}_{[d']} \\ \tilde{\boldsymbol{\xi}}_{d'} \end{pmatrix}.$$

The orthogonal complement $\mathbf{U}$ of given design matrix $\mathbf{Z}$ can be computed by (Haber 1985)

$$\mathbf{U} = \underline{\mathbf{U}} - \mathbf{P}_{\mathbf{Z}}\underline{\mathbf{U}} = \left(\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\right)\underline{\mathbf{U}} \tag{5.38}$$

with any full column rank matrix $\underline{\mathbf{U}}$ and projection matrix $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1})\mathbf{Z}^T$. The orthogonal complement of $\mathbf{Z}_{[d]}$ is denoted by $\mathbf{U}_{[d]}$ and that of $\tilde{\mathbf{Z}}$ by $\tilde{\mathbf{U}}$. From

$$\mathbf{P}_{\tilde{\mathbf{Z}}} = \begin{pmatrix} \mathbf{P}_{\mathbf{Z}_{[d]}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_d \end{pmatrix}$$

follows

$$\tilde{\mathbf{U}} = \mathbf{P}_{\tilde{\mathbf{Z}}} \begin{pmatrix} \underline{\mathbf{U}}_{[d]} \\ \underline{\mathbf{U}}_d \end{pmatrix} = \begin{pmatrix} \mathbf{U}_{[d]} \\ \mathbf{0} \end{pmatrix}$$

assuming that $\mathbf{U}_{[d]}$ was constructed as $\mathbf{U}_{[d]} = (\mathbf{I}_{[d]} - \mathbf{P}_{\mathbf{Z}_{[d]}})\underline{\mathbf{U}}_{[d]}$. The length of the

Lagrange multiplier vector $\tilde{\boldsymbol{\lambda}}$ equals the number of columns of $\tilde{\mathbf{U}}$ and is identical to the length of $\boldsymbol{\lambda}_{[d]}$. The starting values for both Lagrange multipliers are assumed to be equal. Now we partition matrix $\tilde{\mathbf{L}} = \mathbf{L}$ as

$$\frac{\partial \mathbf{L}^T}{\partial \boldsymbol{\xi}} = \begin{pmatrix} \frac{\partial \mathbf{L}_{[d]}^T}{\partial \boldsymbol{\xi}} & \frac{\partial \mathbf{L}_d^T}{\partial \boldsymbol{\xi}} \end{pmatrix} = \begin{pmatrix} \frac{\partial \mathbf{L}_{[d]}^T}{\partial \boldsymbol{\xi}_{[d']}} & \frac{\partial \mathbf{L}_d^T}{\partial \boldsymbol{\xi}_{[d']}} \\ \frac{\partial \mathbf{L}_{[d]}^T}{\partial \boldsymbol{\xi}_{d'}} & \frac{\partial \mathbf{L}_d^T}{\partial \boldsymbol{\xi}_{d'}} \end{pmatrix}.$$

The off diagonal blocks of $\frac{\partial \mathbf{L}}{\partial \boldsymbol{\xi}}$ are zero, because we assume deleting set d' is equivalent to deleting set d. Consequently matrix $\tilde{\mathbf{H}}$ simplifies to

$$\tilde{\mathbf{H}} = \frac{\partial \mathbf{L}^T}{\partial \boldsymbol{\xi}} \tilde{\mathbf{U}} = \begin{pmatrix} \frac{\partial \mathbf{L}_{[d]}^T}{\partial \boldsymbol{\xi}_{[d']}} \mathbf{U}_{[d]} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{H}_{[d]} \\ \mathbf{0} \end{pmatrix}.$$

It also follows that $\tilde{\mathbf{h}} = \tilde{\mathbf{U}}^T \tilde{\mathbf{L}} = \mathbf{U}_{[d]}^T \mathbf{L}_{[d]} = \mathbf{h}_{[d]}$. Matrix $\tilde{\mathbf{S}}$ and vector $\tilde{\mathbf{s}}$ for the "replace=augment" method are partitioned as follows

$$\tilde{\mathbf{S}} = \begin{pmatrix} \mathbf{D}(\mathbf{m}_{[d']}) & \mathbf{H}_{[d]} & \mathbf{0} \\ \mathbf{H}_{[d]}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}(\mathbf{m}_{d'}) \end{pmatrix} \qquad \tilde{\mathbf{s}} = \begin{pmatrix} \mathbf{y}_{[d']} - \mathbf{m}_{[d']} + \mathbf{H}_{[d]} \boldsymbol{\lambda}_{[d]} \\ \mathbf{h}_{[d]} \\ \mathbf{y}_{d'} - \mathbf{m}_{d'} \end{pmatrix}.$$

We conclude from the block-diagonal form of $\tilde{\mathbf{S}}$ and (5.23) that the new iterates of parameters $\boldsymbol{\theta}_{[d]}$ (method 1) and $\tilde{\boldsymbol{\theta}}$ (method 2) are equivalent, that is, the pairs of parameters $(\boldsymbol{\xi}_{[d']}, \boldsymbol{\lambda}_{[d]})$ and $(\tilde{\boldsymbol{\xi}}_{[d']}, \tilde{\boldsymbol{\lambda}})$ are identical. The new iterate $\boldsymbol{\beta}_{[d]}^{new}$ can be directly computed from $\boldsymbol{\xi}_{[d']}$ or $\tilde{\boldsymbol{\xi}}_{[d']}$, because it is apparent from

$$\tilde{\boldsymbol{\beta}} = \mathbf{R}_{\tilde{\mathbf{Z}}} \mathbf{L} = \begin{pmatrix} \mathbf{R}_{\hat{\mathbf{Z}}_{[d]}} \mathbf{L}_{[d]} \\ \mathbf{L}_d \end{pmatrix}$$

that $\tilde{\boldsymbol{\beta}}_{[d]}^{new} = \hat{\boldsymbol{\beta}}_{[d]}^{new}$. Thus, the first and second method are equivalent for each step. We have $\bar{\mathbf{Z}} = \mathbf{Z}_{[d]}$ and can assume that $\bar{\mathbf{U}} = \mathbf{U}_{[d]}$. From

$$\frac{\partial \bar{\mathbf{L}}^T}{\partial \boldsymbol{\xi}} = \begin{pmatrix} \frac{\partial \mathbf{L}_{[d]}^T}{\partial \boldsymbol{\xi}_{[d']}} \\ \frac{\partial \mathbf{L}_{[d]}^T}{\partial \boldsymbol{\xi}_{d'}} \end{pmatrix}$$

follows

$$\bar{\mathbf{H}} = \begin{pmatrix} \mathbf{H}_{[d]} \\ \mathbf{0} \end{pmatrix}.$$

It follows $\bar{\mathbf{S}}^{-1}\bar{\mathbf{s}} = \tilde{\mathbf{S}}^{-1}\tilde{\mathbf{s}}$ and thus (1a) of the theorem.

Equivalence of methods 2 and 3: Now there is no partition of $\boldsymbol{\xi}$. We have

$$\frac{\partial \bar{\mathbf{L}}^T}{\partial \boldsymbol{\xi}} = \frac{\partial \mathbf{L}_{[d]}^T}{\partial \boldsymbol{\xi}}, \quad \frac{\partial \tilde{\mathbf{L}}^T}{\partial \boldsymbol{\xi}} = \begin{pmatrix} \frac{\partial \mathbf{L}_{[d]}^T}{\partial \boldsymbol{\xi}} & \frac{\partial \mathbf{L}_{d}^T}{\partial \boldsymbol{\xi}} \end{pmatrix}$$

and it follows $\bar{\mathbf{H}} = \tilde{\mathbf{H}} = \mathbf{H}_{[d]}$. Therefore: $\bar{\mathbf{S}}^{-1}\bar{\mathbf{s}} = \tilde{\mathbf{S}}^{-1}\tilde{\mathbf{s}}$ with

$$\bar{\mathbf{S}}^{-1}\bar{\mathbf{s}} = \begin{pmatrix} \mathbf{D}(\mathbf{m}) & \mathbf{H}_{[d]} \\ \mathbf{H}_{[d]}^T & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{y} - \mathbf{m} + \mathbf{H}_{[d]}\boldsymbol{\lambda}_{[d]} \\ \mathbf{h}_{[d]} \end{pmatrix}.$$

Identical final solutions for methods 1 and 3: We can expect function $\mathbf{L}(\cdot)$ to have identical values (at least approximate) for methods 1 and 3 at convergence, since both methods are subject to the same model expressed now only in terms of $\mathbf{L}_{[d]}(\cdot)$. Matrix $\mathbf{Z}_{[d]}$ is identical for both methods, therefore, parameter estimates for both methods are computed by $\mathbf{R}_{\mathbf{Z}_{[d]}}\mathbf{L}_{[d]}(\cdot)$. Part (b) of the theorem follows. □

Theorem 5.3.7. *For the deletion of the set d, we have the following one-step approxima-*

tions

$$\begin{aligned}\mathbf{m}_{[d]} \approx (-\mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{H}_{[d]}(\mathbf{H}_{[d]}^T\mathbf{D}^{-1}\mathbf{H}_{[d]})^{-1}\mathbf{H}_{[d]}^T\mathbf{D}^{-1})(\mathbf{v} - \mathbf{m} + \mathbf{H}_{[d]}\mathbf{1}_{[d]}) \\ + \mathbf{D}^{-1}\mathbf{H}_{[d]}(\mathbf{H}_{[d]}^T\mathbf{D}^{-1}\mathbf{H}_{[d]})^{-1}\mathbf{h}_{[d]},\end{aligned} \tag{5.39}$$

and

$$\boldsymbol{\beta}_{[d]} = \mathbf{R}_{\mathbf{Z}_{[d]}}\mathbf{L}_{[d]}(\mathbf{m}_{[d]}).$$

If the deletion of set d' does not correspond to a deletion of any set d, we find the approximations

$$\begin{aligned}\mathbf{m}_{[d']} \approx (-\mathbf{D}_{[d']}^{-1} + \mathbf{D}_{[d']}^{-1}\mathbf{H}_{[d']}(\mathbf{H}_{[d']}^T\mathbf{D}_{[d']}^{-1}\mathbf{H}_{[d']})^{-1}\mathbf{H}_{[d']}^T\mathbf{D}_{[d']}^{-1})(\mathbf{v}_{[d']} - \mathbf{m}_{[d']} + \mathbf{H}_{[d']}\mathbf{1}) \\ + \mathbf{D}_{[d']}^{-1}\mathbf{H}_{[d']}(\mathbf{H}_{[d']}^T\mathbf{D}_{[d']}^{-1}\mathbf{H}_{[d']})^{-1}\mathbf{h}_{[d']}\end{aligned}$$

and

$$\boldsymbol{\beta}_{[d']} = \mathbf{R}_{\mathbf{Z}}\mathbf{L}(\mathbf{m}_{[d']}).$$

A one-step approximation of the Cook distance is obtained with (5.26). The difference in the likelihood ratio test due to the deletion of subset d or subset d' (denoted by d/d') is

$$L^2(\boldsymbol{\beta}) - L^2(\boldsymbol{\beta}_{[d/d']}) = 2\mathbf{y}^T\log(\mathbf{y}(\mathbf{m}_{[d/d']}/\mathbf{m}))$$

Proof. The one-step approximations (5.39) and (5.3.7) follow directly from applying one step of (5.23) and by using (5.24) with $\boldsymbol{\lambda} = \mathbf{1}$. The other deletion diagnostics are functions of $\mathbf{m}_{[d/d']}$ and $\mathbf{m}$. □

Remark 5.3.8. The formulae involve the orthogonal complement $\mathbf{U}$ of $\mathbf{Z}$ which can be computed from (5.38). The matrix $\mathbf{H} = \frac{\partial \mathbf{h}(\mathbf{m})^T}{\partial \mathbf{m}} = \frac{\partial \mathbf{L}^T}{\partial \mathbf{m}}\mathbf{U}$ is of size $(K \cdot 2^J) \times (K - p)$ ($\mathbf{L} \in \mathbb{R}^K$, $\mathbf{m} \in \mathbb{R}^{K \cdot 2^J}$ and $\mathbf{U} \in \mathbb{R}^{K \times (K-p)}$).Consequently $\mathbf{H}_{[d']}$ refers to

the deletion of rows of $\mathbf{H}$ and $\mathbf{H}_{[d]}$ refers to the deletion of d of the K covariate settings, such that K is reduced to $K - |d|$.

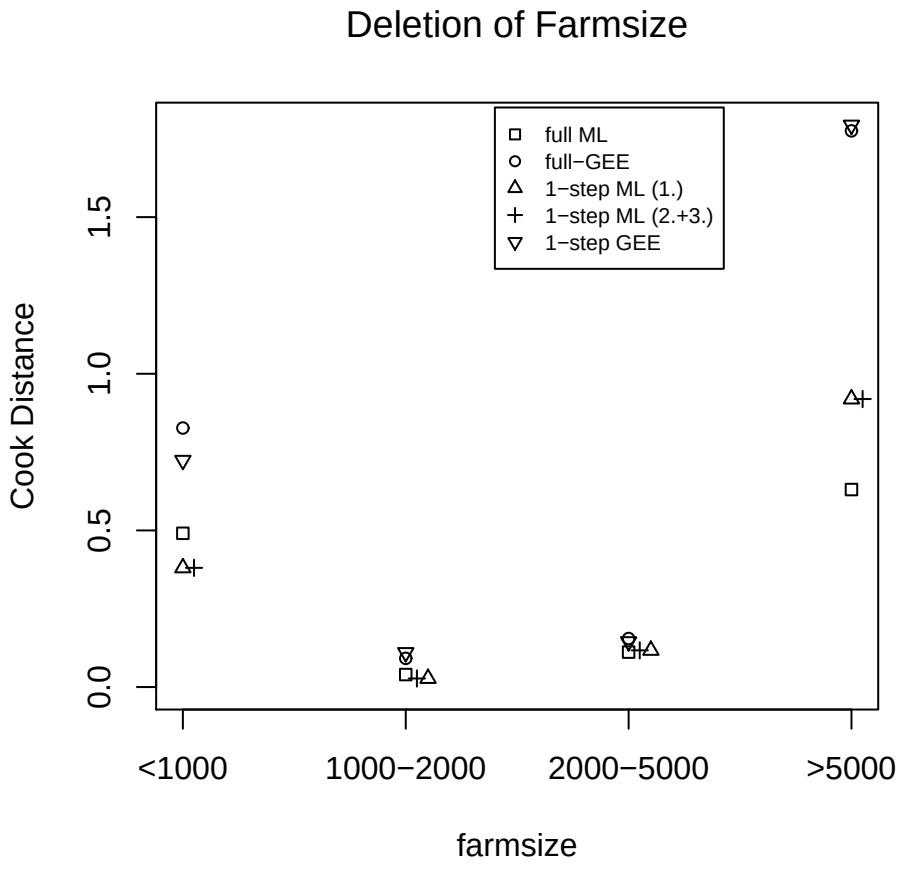
5.4 Example

5.4.1 Deletion of Predictors

For the farmers' data and model "LIN S", we investigate the influence of deleting components of $\mathbf{L}$ and $\mathbf{Z}$ that are in set d, which is the influence of predictors contained in design matrix $\mathbf{Z}$. Figures 5.1 and 5.2 show full solutions and one step approximations for the Cook distance for the farmers' data and model (5.4) with GEE and HLP fitting algorithms deleting farmsize (Figure 5.1) and item×farmsize (Figure 5.2). Deleting education levels does not seem sensible, because the predictors for model "LIN S" do not depend on education. Deleting one level of farmsize can be accomplished by either deleting 2 of the 8 columns of joint observations in Table 5.2 or by deleting $10 = 2 \times 5$ components of function $\mathbf{L}$ which is of length $40 = 8 \times 5$ ($K = 8$, $J = 5$). In contrast, deleting item×farmsize can be achieved by deleting components of $\mathbf{L}$ and $\mathbf{Z}$, but not by deletion of joint observations. The joint observations for the given farmsize level have to be changed in such a way that the multiple responses have the item removed. We conclude, for deleting farmsize, there is a set d' that corresponds to set d, whereas for deleting item×farmsize is no such set d' that is equivalent to d.

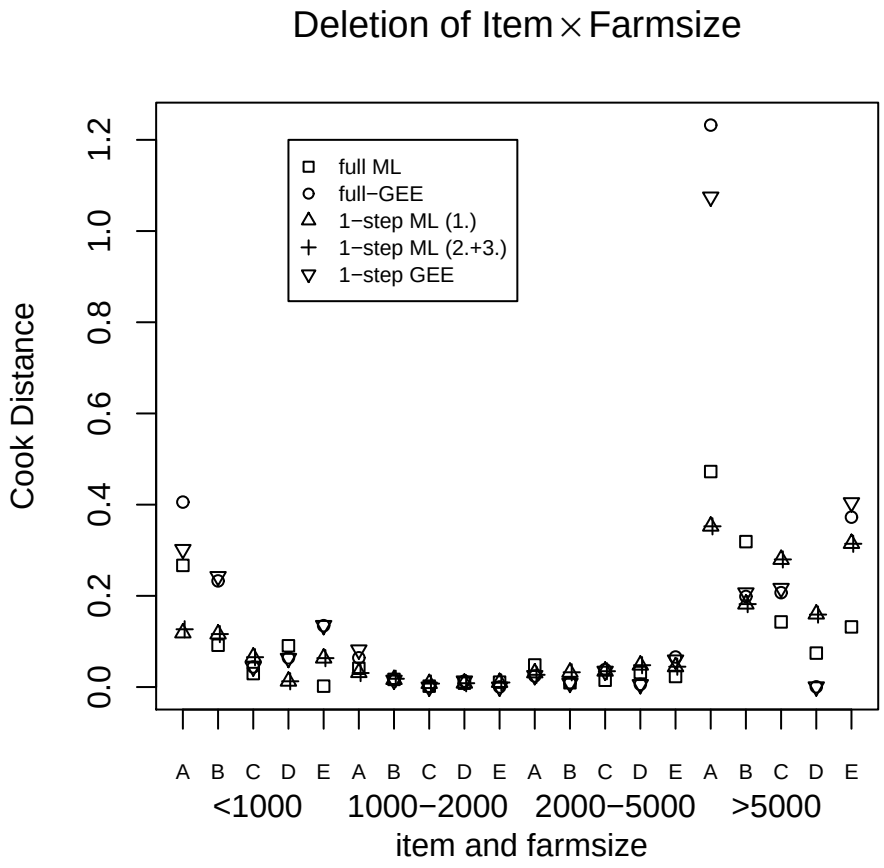
The results confirm that the 3 HLP deletion methods as well as the 2 GEE deletion methods are equivalent. Only one-step approximations of method 1 differ slightly from those of methods 2 and 3, if condition (a) of Theorem 5.3.6 is not fulfilled. However, the difference is negligible, see Figure 5.2. When comparing the Cook distance for GEE and HLP, we can say the following: Both methods tend

Figure 5.1: Cook Distance for model (5.4) and deletion of farm size



to give similar results, as small/large values for HLP will also give small/large values for GEE, however, the exact values differ and tend to vary more for GEE. Generally, messages regarding influence seem similar. For example, for deleting item A and farmsize > 5000, the Cook distance for GEE is around 1.2 and for HLP is only around 0.5. However, both values are relatively large compared to the other values and suggest that observations for this combination of predictors are influential on model "LIN S". Figure 5.1 indicates that farmsize level > 5000 is influential. Figure 5.2 presents a clearer picture showing that item A with farm-

Figure 5.2: Cook Distance for model (5.4) and deletion of item×farmsize



size level > 5000 is most influential and other items with farmsize level > 5000 are probably not influential.

5.4.2 Deletion of Joint Observations

Now we investigate the deletion of joint observations with setting k and outcome j'. The number of those observations is $v_{kj'}$. The farmers' data has $K = 8$ different covariate settings. For each setting k, we have $2^J = 32$ possible outcomes j'. Some of them were not observed; some, however, were recorded multiple times. If

$v_{kj'} > 1$, then it does not seem sensible to delete only one case, reducing $v_{kj'}$ by 1, because the remaining $v_{kj'}-1$ observations still have an influence on the estimates similar to the original $v_{kj'}$ observations. It rather seems plausible to delete all $v_{kj'}$, such that after deletion $v_{kj'} = 0$. One problem remains, some entries of Table 5.2 are relatively large $v_{kj'} = 14$, whereas other nonzero entries are small $v_{kj'} = 1$. We expect the influence of those 14 observations to be larger than the influence of other observations with $v_{kj'} = 1$. Hence, it seems wiser to divide the Cook distance by the number $|\mathbf{v}_{d'}|$ of observations being deleted

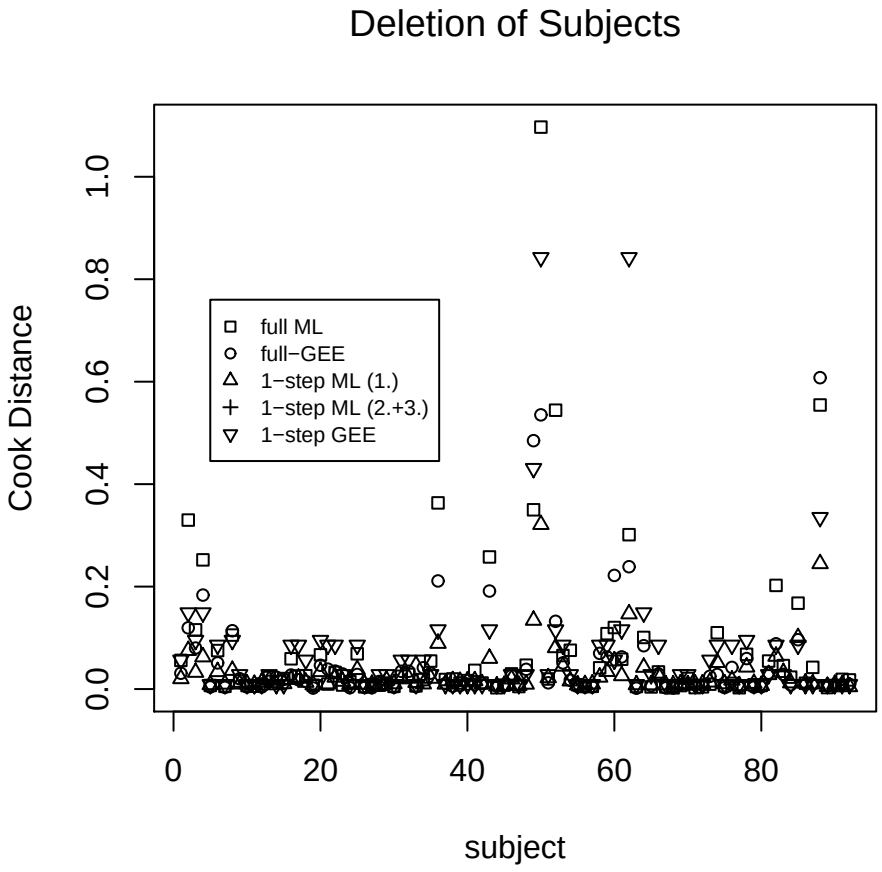
$$CD^s_{d'} = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[d]})^T \text{Cov}(\hat{\boldsymbol{\beta}})^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[d]})/(p|\mathbf{v}_{d'}|). \tag{5.40}$$

Figure 5.3 shows the Cook distance for deleting all responses with outcome j' and setting k, whereas Figure 5.4 shows the standardised Cook distance defined in (5.40). In Figure 5.3, the Cook distance is largest for those observations for which $v_{ij'}$ is largest, for example the highest values are obtained for $v_{ij'} = 14, 10, 10$, which is to be expected and not satisfactory in the detection of influential observations. In contrast, Figure 5.4 shows a much more balanced picture. We can conclude, no observation seems to have a large influence on $\hat{\boldsymbol{\beta}}$.

5.5 Discussion

Both, GEE and HLP (ML) deletion diagnostics have their limitations. GEE is not based on maximum likelihood and should only be applied if HLP diagnostics are not applicable due to either too many zero cell counts or the huge number of multinomial parameters. In particular, for non-grouped observations ($v_{k+} = 1$) and a large number of items, the HLP model methodology seems infeasible because the ratio of nonzero and zero entries is one-to-many.

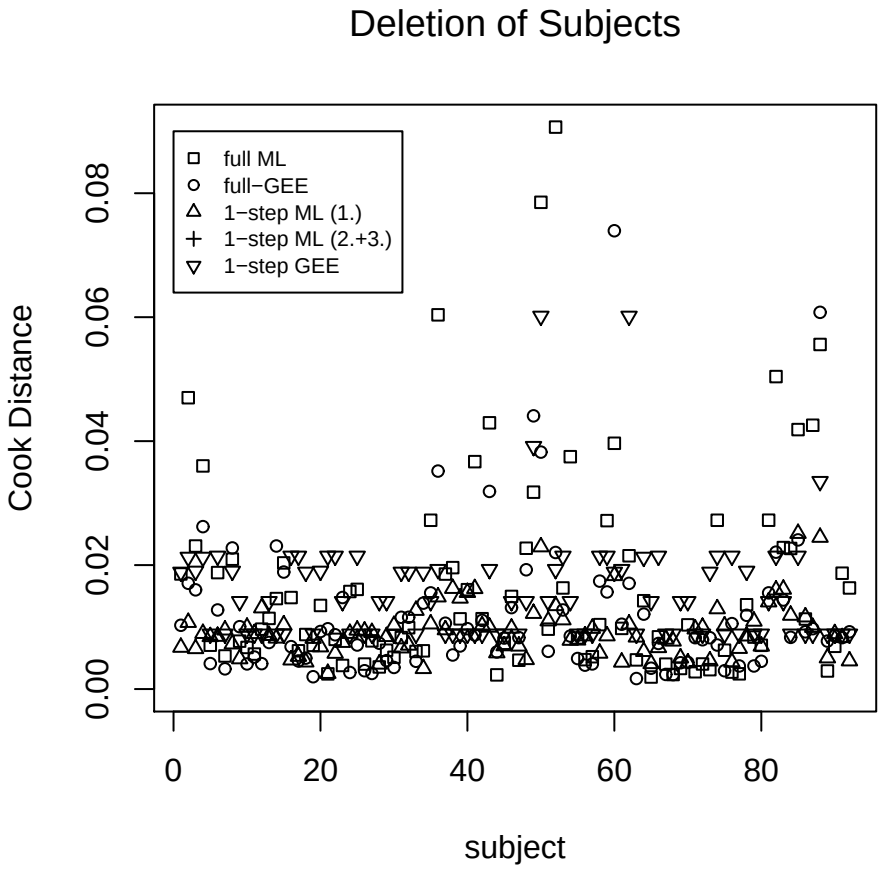
Figure 5.3: Cook Distance for model (5.4) and deletion of subjects



We investigated HLP diagnostics for marginal models for multiple response data, however, the introduced deletion methods do not depend on marginal models only, but are generally applicable for GEE and HLP models. Furthermore, the deletion methods do not depend on GEE and HLP models only, but on the corresponding iteration schemes (5.6) and (5.23) and can also be applied for any other model approach with an identical fitting algorithm.

Generally, the "delete=augment" method is a useful tool in computing deletion diagnostics, because it only requires the manipulation of the design matrix

Figure 5.4: Standardised Cook Distance for model (5.4) and deletion of subjects



and all other quantities can be left unchanged to obtain either full solutions or one-step approximations. Furthermore, for HLP models it is recommended to check whether the deletion of a set d' of (joint) observations is equivalent to deleting a set d referring to the rows of the design matrix. If this is true, the deletion of d is much simpler to handle than that of d' and is to be preferred. Again, we can apply the relative simple "delete=augment" method. The results show that the deletion of a set of predictors d seems more plausible than the deletion of a set d' of single joint observations (not corresponding to deleting a set d of predictors).

This is because the results of Figures 5.3 and 5.4 are more difficult to interpret than those of Figures 5.1 and 5.2. If such a set d' is deleted, then we recommend using the standardised Cook distance (5.40).

Chapter 6

Repeated Multiple Responses

6.1 Introduction

In this chapter, we discuss the modelling of a repeated multiple response variable, a categorical variable for which subjects can select any number of categories on repeated occasions. Multiple responses have been considered in the literature by various authors, e.g. Loughin and Scherer (1998), Agresti and Liu (1999), Agresti and Liu (2001), however, repeated multiple responses have not yet been considered.

Students of a statistics lecture (STAT 291) at the Victoria University of Wellington (New Zealand) were asked by their lecturer, Dr Ivy Liu, to complete a questionnaire on 3 different occasions: 2004, July 2005 and October 2005. They were asked the following questions and to tick the appropriate boxes:

1. "Indicate which of these Wellington bars you have been to" and which of these ticked is your most favourite bar. Any/75 bars could be chosen plus the option "other" bar, where the student was also asked to provide its name.

2. "What type(s) of music do you listen to when you go out to bars? (a) Alternative, (b) Dance, (c) Hip Hop, (d) Karaoke, (e) Pop, (f) Rock, (g) 6os, (h) 7os, (i) 8os, (j) 9os, (k) Other (please specify)."

3. "Do you prefer to dress up to go out to bars? Yes/No"

4. "Do you enjoy playing pool?: Yes/No"

5. "Do you get out to ... ? (a) Socialise with friends, (b) Meet new people, (c) Listen to music, (d) Get drunk, (e) Other (please specify)."

6. "Do you think your choice of bar is affected by advertising? Yes/No"

7. "How many bars would you visit on a night out? (a) $1-2$, (b) $3-4$, (c) $5-6$, (d) 7 or more."

8. "Is a bar's décor usually important to you? For instance, how the place looks. Yes/No"

9. "Is a bars popularity important to you? Yes/No"

10. "How often do you go out to bars? (a) Once a day, (b) Every second day, (c) Once a week, (d) Every second week, (e) Once a month."

11. "Do you drink alcohol? Yes/No"

12. "Do you smoke cigarettes? Yes/No"

13. "Do you work? (a) Yes (full-time or part-time), No"

14. "How long have you lived in Wellington? (a) ≤ 5 months, (b) $6-11$ months, (c) $12-17$ months, (d) $18-23$ months, (e) ≥ 24 months."

Our aim is to model how the choice of the favourite bar is affected and associated by the bars' features and how it depends on the responses to questions (2)-(14) but also on some other fixed covariates such as age, sex, major, ethnicity and type of fees.

Let $y_{ijt} = 1$ if subject $i = 1, \ldots, n$ selects category $j = 1, \ldots, J$ at time point or occasion $t = 1, \ldots, T$ and $y_{ijt} = 0$ otherwise. Let $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \ldots, \mathbf{y}_{iT}^T)^T$ with $\mathbf{y}_{it} = (y_{i1t}, y_{i2t}, \ldots, y_{iJt})^T$ denote the response profile on the J categories and T time points. Note that superscript T denotes the transpose of a vector/matrix and subscript T refers to the number of time points. We regard "Drink Deals", "Pool Table" and "Sports TV" as responses by recording each student's favourite bar at time $t = 1, \ldots, 3$ and by setting $y_{i1t} = 1$, if the student's favourite bar offers "Drink Deals", $y_{i2t} = 1$, if the student's favourite bar is equipped with a "Pool Table" and $y_{i3t} = 1$, if the student's favourite bar also offers some sort of "Sports TV", and $y_{ijt} = 0$ otherwise. Actually, the students only select their favourite bar at occasion t and then, from this univariate response and from the bar's features we obtain a multivariate binary sequence $\mathbf{y}_{it}$, which we regard as multiple responses.

For example: The first student ticked "Zebos" as his favourite bar in 2004 (t=1) and "Kitty" in July 2005 (t=2), whereas in Oct 2005 (t=3) his response was not available (NA). The third student's favourite bar was the "Occidental" at all 3 times. The 10th student responded only twice (t=1,2) with "Havana", unfortunately, the features of "Havana" were not recorded and repeated multiple responses were all set to "not available" (NA). The bar "Kitty" offers all three features "Drink Deals","Pool Table" and "Sports TV". The bar "Zebos" only offers "Drink Deals" and "Pool Table", whereas "Occidental" can only offer "Sports TV". We obtain the following repeated multiple responses for those students:

$\mathbf{y}_1 = (\mathbf{y}_{1,1}^T, \ldots, \mathbf{y}_{1,3}^T)^T = (1, 1, 0, 1, 1, 1, NA, NA, NA)^T$, $\mathbf{y}_3 = (\mathbf{y}_{3,1}^T, \ldots, \mathbf{y}_{3,3}^T)^T = (0, 0, 1, 0, 0, 1, 0, 0, 1)^T$ and $\mathbf{y}_{10} = (NA, \ldots, NA)^T$. In the following, we will refer to this example as STAT 291 data.

Similar to Agresti and Liu (2001) who considered "modelling strategies for multiple response data", this chapter considers several strategies for modelling repeated multiple response data using existing methods.

The next section introduces a marginal model approach for repeated multiple responses. In the next two sections, we discuss the ML (Section 6.3) and GEE (Section 6.4) fitting approaches for the marginal models. In Section 6.4, we also consider possible correlation structures and propose a groupwise correlation estimation method, yielding more efficient parameter estimates if the correlation structure is indeed different for different groups, which is confirmed by a simulation study. Section 6.5 considers generalised linear mixed models (GLMM) with normal random effects as an alternative to the marginal model approach. Section 6.6 discusses parameter estimation results for the STAT 291 data and the final section compares strategies, shows interconnections between them and gives some recommendations.

6.2 Marginal Modelling

We use similar notations as in Section 5.2 on page 149, where we introduced the marginal modelling approach for multiple responses. The vector $\mathbf{y}_i$ contains the T multiple response variables $\mathbf{y}_{it} \in \mathbb{R}^J$ for subject i and occasion t forming a variable of length $J \times T$. The $J \times T$ components of $\mathbf{y}_i$ are also referred to as items. Each subject's response profile $(y_{i11}, \ldots, y_{iJT})$ contributes to one of the $2^{J \cdot T}$ cells in a contingency table cross classifying the items. We assume that observations

for such tables are independent and follow a multinomial distribution with $2^{J \cdot T}$ possible outcomes. For covariate setting $k = 1, \ldots, K$, let the number of multinomial (or joint) observations with outcome j' be denoted by $v_{kj'}, j' = 1, \ldots, 2^{J \cdot T}$, where j' refers to one of the outcomes of the form $(j'_{1,1}, \ldots, j'_{J,T}), j'_{j,t} \in \{0, 1\}$. This is the same index j', we introduced in Section 5.2.3 on page 157, the only difference is that the binary sequences j' are now of length $J \times T$. Table 5.2 on page 158 shows responses j' of length $J = 5$ for the farmers' data. In contrast to Chapter 5 and the farmers' data, the observations of the STAT 291 data all have a unique covariate setting, such that $v_{k+} = 1$, or in other words, the ith subject has covariate setting $i = 1, \ldots, n(= K)$. Similarly denote the multinomial (or joint) probabilities for setting k by $\tau_{kj'}, j' = 1, \ldots, 2^{J \cdot T}$. In the following, we use index i to refer to the ith subject but also to the ith covariate setting. Also let π_{ijt} denote the (marginal) probability of a positive response for observation i, category j and occasion t, which can be computed by $\pi_{ijt} = \sum_{\{(j'_{1,1}, \ldots, j'_{J,T}): j'_{j,t}=1\}} \tau_{ij't}$. Note that $0 \leq \sum_{j,t} \pi_{ijt} \leq J \cdot T$. Let $\boldsymbol{\pi}_i$ denote the vector containing the marginal probabilities, similarly $\mathbf{v}_i$ and $\boldsymbol{\tau}_i$. The marginal probabilities can be computed from the joint probabilities by $\boldsymbol{\pi}_i = \mathbf{B}\boldsymbol{\tau}_i$ with matrix $\mathbf{B}$ containing only 0s and 1s, see Section 5.2.3 on page 157 for more details.

For each subject i, let a column vector of fixed covariates $\mathbf{x}_{i0}$ and time-dependent covariates $\mathbf{x}_{it}, t = 1, \ldots, T$ (also row vectors) be given and let $\mathbf{x}_i = (\mathbf{x}_{i0}^T, \mathbf{x}_{i1}^T, \ldots, \mathbf{x}_{iT}^T)^T$ be the vector containing all covariates. Now we model the probabilities π_{ijt} in terms of the covariates $\mathbf{x}_{it}, t \geq 0$ by

$$g_j(\pi_{ijt}) = \alpha_{jt} + \mathbf{x}_{i0}^T \boldsymbol{\beta}_{0j} + \mathbf{x}_{it}^T \boldsymbol{\beta}_{tj} = \mathbf{z}_{ijt}^T \boldsymbol{\beta}_{jt} = \eta_{ijt}, \tag{6.1}$$

where g_j is the jth link function, η_{ijt} the linear predictor, α_j the j-th intercept

parameter, $\mathbf{z}_{ijt}$ the corresponding vector of the design matrix depending on $\mathbf{x}_i$, or in vector form

$$\mathbf{g}(\boldsymbol{\pi}_i) = \mathbf{Z}_i\boldsymbol{\beta} = \boldsymbol{\eta}_i,$$

with $\mathbf{g} = (g_1, g_2, \ldots, g_J, \ldots, g_1, g_2, \ldots, g_J)^T$, $\mathbf{Z}_i = \text{Diag}(\mathbf{z}_{i11}^T, \ldots, \mathbf{z}_{iJ1}^T, \ldots, \mathbf{z}_{i1T}^T, \ldots, \mathbf{z}_{iJT}^T)$, $\boldsymbol{\beta} := (\alpha_1, \ldots, \alpha_J, \boldsymbol{\beta}_{01}^T, \ldots, \boldsymbol{\beta}_{JT}^T)^T$, $\boldsymbol{\pi}_i = (\pi_{i11}, \ldots, \pi_{iJ1}, \ldots, \pi_{i1T}, \ldots, \pi_{iJT})^T$, $\boldsymbol{\eta}_i = (\eta_{i11}, \ldots, \eta_{iJ1}, \ldots, \eta_{i1T}, \ldots, \eta_{iJT})^T$.

Assume a common effect $\beta_j = \beta_{1j} = \cdots = \beta_{Tj}$ and a logit link then model (6.1) becomes

$$\log\left(\frac{\pi_{ijt}}{1-\pi_{ijt}}\right) = \alpha_j + \mathbf{x}_{i0}^T\boldsymbol{\beta}_{0j} + \mathbf{x}_{it}^T\boldsymbol{\beta}_j. \tag{6.2}$$

For fixed j and t, the model is the ordinary logit model, where the effect varies according to outcome category j. For certain data, one might also consider the same effect over all J categories.

6.3 Maximum Likelihood Estimation

Assuming independence between all items would make the fitting quite simple by using ordinary software for generalised linear models (GLM) (McCullagh and Nelder 1989). However, the more efficient way is fitting the J models simultaneously. Previously, we introduced marginal and multinomial (joint) probabilities and observations. Define the multinomial expected cell counts by $m_{ij} = v_{i+}\tau_{ij}$ or equivalently in vector form $\mathbf{m}_i := v_{i+} \cdot \boldsymbol{\tau}_i$.

Maximum likelihood (ML) estimates are obtained by maximising the log-likelihood kernel

$$\sum_{i=1}^{n}\sum_{j'=1}^{2^{JT}} v_{ij'} \log m_{ij'} = \sum_{i=1}^{n} \mathbf{v}_i^T \log \mathbf{m}_i \tag{6.3}$$

subject to model (6.1). Lang (2005) introduced homogeneous linear predictor

(HLP) models which have the form

$$\mathbf{L}(\mathbf{m}_i) = \mathbf{Z}_i\boldsymbol{\beta}$$

with homogenous link function $\mathbf{L}$. The approach formulates the ML estimation problem as a constrained maximisation problem, where the model is formulated as a system of constraints. Model (6.1), with a sufficiently smooth link function $g_j(\cdot)$, is a HLP model, e.g. logit link or probit link. Fitting of HLP models was discussed in Subsection 5.2.3 on page 157.

Fitzmaurice and Laird (1993) proposed another ML method to obtain parameter estimates. They derived likelihood equations for the mean response and association parameters by expressing the likelihood in terms of the model parameters. However, given the estimates, these equations only determine the first and second order moments of the joint distribution, but the full joint distribution (including the higher order moments) cannot be determined. They circumvented this problem by applying the IPF algorithm for given parameter iterates to get a solution for the joint distribution for each step of the fitting algorithm.

However, ML-estimation has some severe drawbacks for our type of data. For our example, we have $J = T = 3$ resulting in $2^{JT} = 2^9 = 512$ joint probabilities for each of the 122 students. Although some students will be deleted due to NA entries, the amount of computer memory required is still quite large and makes the method almost infeasible despite quite small values for J and T. In this instance, the standard, modern computers available to us failed to give parameter estimates using the HLP fitting algorithm by running out of memory. Fitzmaurice and Laird (1993)'s method is even more complex, because it also requires the application of the IPF algorithm in each step.

Another problem is that the ML method requires (theoretically) non-zero cell counts. For non-grouped observations there is only 1 observation per table with 2^{JT} cells. For instance, for $J = 2$ and $T = 1$, there is only one out of $2^{JT} = 4$ cells that are nonzero. The ratio becomes even worse for larger J and T . Each joint table i with observations $v_{ij'}$ represents a sample, but one observation can be hardly considered as such. A very small constant (e.g. 10^{-5}) is usually added to those zero cell counts to avoid convergence problems. However, the huge number of those zero cell counts for repeated multiple responses will lead to severe convergence problems. Unless J and T are very small (the product JT is ≤ 6) and the observations are grouped (like the farmers' data, see Table 5.1 on page 148), we do not recommend ML estimation and it will not be considered here.

Robins, Rotnitzky and Zhao (1995, p. 106) point out, that "ML methods can be sensitive to model misspecification, because they implicitly impute the missing data from their conditional distribution given the observed data". Hence, our concerns do not only arise from the huge number of zero cell counts, but also from missing data. In the next section, we discuss a quasi-likelihood approach.

6.4 Generalised Estimation Equations

6.4.1 Introduction

As mentioned earlier, when wrongly assuming independence between the $J \times T$ items, the generalised linear models (GLM) (McCullagh and Nelder 1989) methodology can be easily applied yielding ML estimates. However, more efficient parameter estimates can be obtained by the generalised estimation equation (GEE) method (Liang and Zeger 1986), where marginal models are fitted simultaneously and a chosen correlation structure is incorporated, which is an extension of the

quasi-likelihood method (Wedderburn 1974) for multivariate data.

Let $\text{Var}(\mathbf{y}_i) = \mathbf{f}_i \cdot \phi^{-1}$ with variance function $\mathbf{f}_i = \mathbf{f}(\boldsymbol{\pi}_i) = \boldsymbol{\pi}_i(\mathbf{1}_{JT} - \boldsymbol{\pi}_i)$, where $\mathbf{1}_{JT}$ is a vector of length $J \cdot T$, and scale or dispersion parameter ϕ. In the common GEE terminology $\boldsymbol{\mu}_i$ is used instead of $\boldsymbol{\pi}_i$ and the observations $\mathbf{y}_i$ are referred to as clusters with varying cluster length $J_i \leq JT$. Note for model (6.1), $\boldsymbol{\pi}_i$ is identical to the mean $\boldsymbol{\mu}_i$. Suppose model (6.1) is true, then the GEE estimates are obtained by computing the root of the generalised estimation (or quasi-score) equations

$$\sum_{i=1}^{n} \mathbf{M}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i = 0, \tag{6.4}$$

which were introduced in the last chapter by equation (5.5) on page 151 with $\mathbf{M}_i = \partial\boldsymbol{\pi}_i/\partial\boldsymbol{\beta}$, $\mathbf{V}_i = \mathbf{A}_i\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i$ and $\mathbf{r}_i = (\mathbf{y}_i - \boldsymbol{\pi}_i)$. Now $\mathbf{M}_i$ is a $JT \times p$ matrix (p number of parameters), $\mathbf{A}_i = \sqrt{\mathbf{f}_i}$ is a $JT \times JT$ diagonal matrix and $\mathbf{R}_i(\boldsymbol{\alpha})$ is the $JT \times JT$ correlation matrix for observation (or cluster) i ($i = 1, \ldots, n$) depending on correlation parameter(s) $\boldsymbol{\alpha}$.

If the correlation is unknown, they must be estimated consistently for every iterate, for example by using the method of moments suggested by Liang and Zeger (1986). For further details of GEE, the choice of the working correlation and correlation parameter estimation, see Subsection 5.2.2 on page 150.

6.4.2 Correlation Structure

In this subsection, we consider specific choices of the correlation structure $\mathbf{R}_i(\boldsymbol{\alpha})$ for multiple response data and repeated multiple response data. The choice of the working correlation is important, because it determines the estimates and their variances. We also propose a new groupwise method potentially yielding more efficient parameter estimates.

The naive variance will be a good estimate, if $\sum_{i=1}^{n} J_i$ is large and if the correlation is correctly specified. On the other hand, the robust estimate will be good if n, the number of clusters, is large (Lawal 2003). For example, for $n < 25$, the robust variance does not provide a good estimate and the correlation structure should be carefully chosen to make use of the naive variance. Choosing a good correlation structure is essential to obtain good variance estimates, and also in obtaining more efficient parameter estimates for $\hat{\boldsymbol{\beta}}$, e.g. see simulation study in Liang and Zeger (1986).

Let us denote the correlation structure $\mathbf{R}_i = \text{Corr}(\mathbf{y}_i)$ by

$$\mathbf{R}_i = \begin{pmatrix} \mathbf{R}_{i11} & \mathbf{R}_{i12} & \cdots & \mathbf{R}_{i1T} \\ \mathbf{R}_{i12} & \mathbf{R}_{i22} & \cdots & \mathbf{R}_{i2T} \\ \vdots & \vdots & & \vdots \\ \mathbf{R}_{i1T} & \mathbf{R}_{i2T} & \cdots & \mathbf{R}_{iTT} \end{pmatrix},$$

where the indices t_1 and t_2 of $\mathbf{R}_{it_1t_2} \in \mathbb{R}^{J\times J}$ refer to the occasions. The submatrices $\mathbf{R}_{itt}$ and $\mathbf{R}_{it_1t_2}$ have the form:

$$\mathbf{R}_{itt} = \begin{pmatrix} 1 & R_{itt,12} & \cdots & R_{itt,1J} \\ R_{itt,12} & 1 & \cdots & R_{itt,2J} \\ \vdots & \vdots & & \vdots \\ R_{itt,1J} & R_{itt,2J} & \cdots & 1 \end{pmatrix} = (R_{itt,j_1j_1})_{j_1,j_2=1}^{J}$$

$$\mathbf{R}_{it_1t_2} = \begin{pmatrix} R_{it_1t_2,11} & R_{it_1t_2,12} & \cdots & R_{it_1t_2,1J} \\ R_{it_1t_2,21} & R_{it_1t_2,22} & \cdots & R_{it_1t_2,2J} \\ \vdots & \vdots & & \vdots \\ R_{it_1t_2,J1} & R_{it_1t_2,J2} & \cdots & R_{it_1t_2,JJ} \end{pmatrix} = (R_{it_1t_2,j_1j_1})_{j_1,j_2=1}^{J}$$

Note, generally matrix $\mathbf{R}_{it_1t_2}$ is not symmetric, but $\mathbf{R}_{itt}$ is.

Non-Repeated (Standard) Multiple Responses

First we consider non-repeated multiple response data ($T = 1$), later we continue with the general case ($T > 1$). Note that for $T = 1$, the matrix $\mathbf{R}_i$ reduces to $\mathbf{R}_{i11}$ and we omit index t referring to the occasions. As outlined in Subsection (5.2.2) on 150, Liang and Zeger (1986) considered the following correlation structures:

- independence: $R_{i,j_1j_2} = 0$ for all $j_1 \neq j_2$ (0 parameter)
- exchangeable: $R_{i,j_1j_2} = \alpha$ for all $j_1 \neq j_2$ (1 parameter)
- $(J-1)$-dependence: $R_{i,j_1j_2} = \alpha_{|j_1-j_2|}$ ($J-1$ parameters)
- unstructured: totally unspecified $R_{i,j_1j_2} = \alpha_{j_1,j_2}$ ($\frac{1}{2}J(J-1)$ parameters)

and estimated the parameters by the method of moments. The structure $(J-1)$-dependence can also be replaced by m-dependence ($m \leq (J-1)$), which is defined as $R_{i,j_1j_2} = \alpha_{|j_1-j_2|}$ for $|j_1 - j_2| \leq m$ and $R_{i,j_1j_2} = 0$ for $|j_1 - j_2| > m$. That is, two observations taken at time points t_1 and t_2 for an individual always have the same correlation provided $|t_1 - t_2|$ is the same. Another option is an autoregressive correlation (AR) structure which indicates that two observations taken close together in time for an individual tend to be more highly correlated than two observations taken further apart in time from the same individual. Formally, $R_{i,j_1j_2} = \alpha^{|j_1-j_2|}$. We consider five structures (increasing order in number of parameters): independence, exchangeable, autoregressive, m-dependence and unstructured. Given any structure, the correlation is assumed to be equal for all observations. The index i of $\mathbf{R}_i$ only stands for different cluster lengths.

Repeated Multiple Responses

Let us now consider repeated multiple responses ($T > 1$). We can apply the same correlation structures to $\mathbf{R}_i$ as we did before for $\mathbf{R}_{i11}$. However, we would not distinguish between occasions and items. Consequently it seems wiser to consider different structures for the submatrices of $\mathbf{R}_i$. First, we consider the submatrices $\mathbf{R}_{itt}$. Every submatrix $\mathbf{R}_{itt}$ can have the structures independence, exchangeable, autoregressive, m-dependence and unstructured, as we considered for standard (non-repeated) multiple response data. We can also consider similar structures for the off-diagonal matrices $\mathbf{R}_{it_1t_2}(\equiv \mathbf{R}_{it_2t_1})$ with $t_1 \neq t_2$, however, generally the diagonal elements of these matrices do not equal one ($R_{it_1t_2,jj} \neq 1$) and we also do not have symmetry ($R_{it_1t_2,j_1j_2} \neq R_{it_1t_2,j_2j_1}$ for $j_1 \neq j_2$ and $t_1 \neq t_2$). Let us consider the following correlation structures for $\mathbf{R}_{it_1t_2}$

- independence: $R_{it_1t_2,j_1j_2} = 0 \; \forall j_1, j_2 = 1, \ldots, J$ (0 parameter)
- exchangeable: $R_{it_1t_2,j_1j_2} = \alpha \; \forall j_1, j_2 = 1, \ldots, J$ (1 parameter)
- autoregressive : $R_{it_1t_2,j_1j_2} = \alpha^{|j_1-j_2|+1} \; \forall j_1, j_2 = 1, \ldots, J$ (1 parameter)
- m-dependence: $R_{it_1t_2,j_1j_2} = \alpha_{|j_1-j_2|+1} \; \forall |j_1 - j_2| = 1, \ldots, m$, $R_{it_1t_2,j_1j_2} = 0$ otherwise (m parameters)
- unstructured (items): totally unspecified (J^2 parameters).

If we use different structures for $\mathbf{R}_{itt}$ and $\mathbf{R}_{it_1t_2}$, we consider two simple options: One might assume a common structure for all submatrices $\mathbf{R}_{i11} = \cdots = \mathbf{R}_{iTT}$ or different structures for different occasions $\mathbf{R}_{i11} \neq \cdots \neq \mathbf{R}_{iTT}$ (similarly $\mathbf{R}_{it_1t_2}$). We will refer to these as *common* and *different*.

Let us now assume the structures for $\mathbf{R}_{itt}$ and $\mathbf{R}_{it_1t_2}$ are not independent, such that $\mathbf{R}_{itt}$ is a sub-case of $\mathbf{R}_{it_1t_2}$. For given time points t_1 and t_2 ($t_1, t_2 = 1, \ldots, T$), we

consider the same structures (independence, exchangeable, autoregressive, m-dependence and unstructured) for $R_{j_1j_2,t_1t_2}$ as we did before for the submatrices $\mathbf{R}_{itt}$ and $\mathbf{R}_{it_1t_2}$. We denote such a structure by "structure (item)" to underline that the structure refers to the items j_1 and j_2 for any given occasions t_1 and t_2.

For longitudinal data, the structures exchangeable, m-dependence and autoregressive are often used to describe the dependence over time. Now we consider the following correlation structures over time for given items j_1 and j_2

- exchangeable (time): $R_{it_1t_2,j_1j_2} = \alpha_{j_1j_2}$ (1 parameter)
- autoregressive (time): $R_{it_1t_2,j_1j_2} = \alpha_{j_1j_2}^{|t_1-t_2|}$ (1 parameter)
- m-dependence (time): $R_{it_1t_2,j_1j_2} = \alpha_{|t_1-t_2|,j_1j_2}$ for $|j_1 - j_2| \leq m$ otherwise $R_{it_1t_2,j_1j_2} = 0$ (m parameters for $j_1 \neq j_2$ respectively. $m-1$ parameters for $j_1 = j_2$)
- unstructured (time): $R_{it_1t_2,j_1j_2} = \alpha_{t_1t_2,j_1j_2}$ ($T(T-1)/2$ respectively. T^2 parameters).

Note some of the following inter-relations: exchangeable (time) and exchangeable (items) is equivalent to exchangeable for the whole matrix $\mathbf{R}_i$, exchangeable and different for both $\mathbf{R}_{itt}$ and $\mathbf{R}_{it_1t_2}$ is equivalent to exchangeable (items) and unstructured (time), and unstructured (time) and unstructured (items) is equivalent to assuming unstructured for the whole matrix $\mathbf{R}_i$.

We believe the second approach, combining the structures for $\mathbf{R}_{itt}$ and $\mathbf{R}_{it_1t_2}$ by assuming conditional structures for items given time points and for time-points given items, with typical time dependence structures, seems a better approach than the first one, which considers separate structures for the submatrices $\mathbf{R}_{itt}$ and $\mathbf{R}_{it_1t_2}$. In particular, the higher the number of time-points is, the more appropriate the second approach becomes.

Remark 6.4.1. Typical time-dependence structures, such as autoregressive and m-dependence, are usually applied to different time-points for one variable. Here items are dependent variables and it is appropriate to apply such time-dependence structures to different time points for each item. The second approach addressed this issue by considering different structures over time and items. Within one time point and different items, it seems inappropriate to apply such time-dependence structures, since the structure between items seems rather arbitrary ("unstructured"). Therefore the consideration of time-dependence structures for the first approach seems inappropriate. One could also re-order items and time points, such that time points of one item are next to each other. In such a way, we could consider time-dependence structures for the sub-matrices of $\mathbf{R}_i$. We could call this the third approach. However even items can be closely related and time-dependence structures can be appropriate in some circumstances.

Groupwise Correlation Estimation

The correlation parameters $\boldsymbol{\alpha}$ can be estimated by the method of moments (Liang and Zeger 1986), see Section 5.2.2 on page 150 for details. However, they assume the correlation structure to be equal for all observations. This assumption is practical in terms of simplicity, but quite unrealistic. Let us assume a second model for the correlation parameters κ_{i,j_1j_2} specified by

$$\mathbf{h}(\boldsymbol{\kappa}_i) = \mathbf{Z}_i^J \boldsymbol{\alpha}$$

with the vector valued link function $\mathbf{h}(\cdot)$, design matrix $\mathbf{Z}_i^J$ depending on the ith subject covariates $\mathbf{x}_i$, and parameter vector $\boldsymbol{\kappa}_i$ comprising of parameters κ_{i,j_1j_2}. Prentice (1988) suggested estimating $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ as the root of two sets of GEE. The

first set of GEE is given by formula (5.5) on page 151, that is

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}(\mathbf{y}_i)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0,$$

and the second by

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\kappa}_i^T}{\partial \boldsymbol{\alpha}} \mathbf{V}(\mathbf{w}_i)^{-1}(\mathbf{w}_i - \boldsymbol{\kappa}_i) = 0, \tag{6.5}$$

where $\mathbf{w}_i$ is the corresponding vector of sample correlations. Matrix $\mathbf{V}(\mathbf{y}_i) = \mathbf{V}_i$ is the same covariance matrix defined previously; it only uses arguments $\mathbf{y}_i$ and $\mathbf{w}_i$ to refer to the working covariances of the observations $\mathbf{y}_i$ and the empirical correlations $\mathbf{w}_i$. The second set of GEE has the same form as the first set, replacing only the quantities of the mean response model by those of the correlation model. The two sets of GEE can also be written as

$$\sum_{i=1}^{n} \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\kappa}_i^T}{\partial \boldsymbol{\alpha}} \end{pmatrix} \begin{pmatrix} \mathbf{V}(\mathbf{y}_i) & \mathbf{0} \\ \mathbf{0} & \mathbf{V}(\mathbf{w}_i) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{y}_i - \boldsymbol{\mu}_i \\ \mathbf{w}_i - \boldsymbol{\kappa}_i \end{pmatrix} = 0. \tag{6.6}$$

Zhao and Prentice (1990) introduced the following set of GEE, also called GEE2, to estimate jointly $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$

$$\sum_{i=1}^{n} \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\alpha}} \\ \frac{\partial \boldsymbol{\kappa}_i^T}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\kappa}_i^T}{\partial \boldsymbol{\alpha}} \end{pmatrix} \begin{pmatrix} \mathbf{V}(\mathbf{y}_i) & \mathbf{V}(\mathbf{y}_i, \mathbf{w}_i) \\ \mathbf{V}(\mathbf{y}_i, \mathbf{w}_i) & \mathbf{V}(\mathbf{w}_i) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{y}_i - \boldsymbol{\mu}_i \\ \mathbf{w}_i - \boldsymbol{\kappa}_i \end{pmatrix} = 0. \tag{6.7}$$

It is obvious, that equation (6.6) treats observations ($\mathbf{y}_i$ and $\mathbf{w}_i$) and models as independent, in contrast to equation (6.7), which uses information about the mutual dependence of both models and observations. For Prentice's approach, also called GEE1, the estimation of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ can be obtained by finding roots for both sets of GEE jointly but also separately. GEE1 yields consistent pa-

rameter estimates $\hat{\boldsymbol{\beta}}$, even if the correlation model is wrongly specified. In contrast, GEE2 does not yield consistent estimates for $\boldsymbol{\beta}$ given a wrongly specified correlation model. Also, GEE2 only provides more efficient estimates than GEE1 if the correlation model is indeed correct.

In reality, we are more interested in the mean response model parameters $\boldsymbol{\beta}$. The association parameters, such as the correlation, are often considered as nuisance parameters. Therefore, we think GEE2 is not a good method, because there is too much uncertainty in the correlation model. Let us take a closer look at the various correlation structures considered in this subsection. Although we have only considered a limited range of structures (correlation models), we are still very uncertain which of those models might be the correct one. Therefore, Prentice's approach seems better, because we do not need the correlation model to be correct to yield consistent estimates $\hat{\boldsymbol{\beta}}$. However, firstly, it needs a second set of GEE, which generally must be solved iteratively, and secondly, if $J \cdot T$ is large and a more complicated structure is chosen, the number of parameters $\boldsymbol{\alpha}$ is large and will automatically result in more convergence problems. With Liang and Zeger's procedure we can estimate the correlation structure in each step directly for the given iterates of $\hat{\boldsymbol{\beta}}^{new}$ without any iterative method. We presented in Subsection 5.2.2 on page 150 formulae for the estimation of the correlation parameters $\boldsymbol{\alpha}$ for several popular structures.

Now assume the simple correlation model that the correlation does not vary for every subject, but only varies for different groups. In the following, we consider a quite simple alternative to GEE1 for the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Assume a finite number G of groups is given, otherwise partition data into groups. Let the number of clusters for group g ($g = 1, \ldots, G$) be denoted by n_g with $\sum_{g=1}^{G} n_g = n$ and assume $\lim_{n\to\infty} n/n_g = a_g > 0$. Now we extend Theorem

5.2.1:

Theorem 6.4.2 ("groupwise method"). *Under mild regularity conditions and given that :*

1. $\hat{\boldsymbol{\alpha}}_g$ *is* ${n_g}^{1/2}$ *consistent given* $\boldsymbol{\beta}$ *and* ϕ *for* $g = 1, \ldots, G$
2. $\hat{\phi}$ *is* $n^{1/2}$ *consistent given* $\boldsymbol{\beta}$,
3. $|\partial\hat{\boldsymbol{\alpha}}_g/\partial\phi|$ *is* $O_p(1)$

then $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ *is asymptotically multivariate Gaussian with zero mean and variance*

$$\lim_{n\to\infty} n\mathbf{J}_1^{-1}\mathbf{J}_2\mathbf{J}_1^{-1}$$

where

$$\mathbf{J}_1 = \sum_{i=1}^{n} \mathbf{M}_i^T\mathbf{V}_i^{-1}\mathbf{M}_i \text{ and } \mathbf{J}_2 = \sum_{i=1}^{n} \mathbf{M}_i^T\mathbf{V}_i^{-1}Cov(\mathbf{y}_i)\mathbf{V}_i^{-1}\mathbf{M}_i.$$

Proof. Liang and Zeger (1986) proved Theorem 5.2.1 on page 152. The only difference between Theorems 5.2.1 and 6.4.2 is condition (1.) and $\mathbf{V}_i$. In Theorem 5.2.1 index i of $\mathbf{R}_i$ only refers to possible different cluster lengths but the correlation itself is assumed to be equal for all observations i. In contrast, in Theorem 6.4.2 matrix $\mathbf{R}_i$ stands for different cluster lengths but also stands for different correlations depending on which group g observation i belongs to. Liang and Zeger (1986) use the following lines to prove theorem 5.2.1:
Write $\boldsymbol{\alpha}^*(\boldsymbol{\beta}) = \hat{\boldsymbol{\alpha}}\{\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})\}$ and under some regularity condition $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ can be approximated by

$$\left[\sum_{i=1}^{n} -\frac{\delta}{\delta\boldsymbol{\beta}}\mathbf{U}_i\{\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta})\}/n\right]^{-1}\left[\sum_{i=1}^{n}\mathbf{U}_i\{\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta})\}/n^{1/2}\right],$$

where

$$\frac{\delta \mathbf{U}_i\{\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta})\}}{\delta \boldsymbol{\beta}} = \frac{\partial \mathbf{U}_i\{\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}} + \frac{\partial \mathbf{U}_i\{\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta})\}}{\partial \boldsymbol{\alpha}^*} \frac{\partial \boldsymbol{\alpha}^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$
$$= A_i + B_i C.$$

Let $\boldsymbol{\beta}$ be fixed and Taylor series expansion gives

$$\frac{\sum_{i=1}^n \mathbf{U}_i\{\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta})\}}{n^{1/2}} = \frac{\sum \mathbf{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})}{n^{1/2}} + \frac{\sum \partial \mathbf{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha}}{n} n^{1/2}(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}) + o_p(1) \tag{6.8}$$

$$= A^* + B^* C^* + o_p(1).$$

Now $B^* = o_p(1)$, since $\partial \mathbf{U}_i/\partial \boldsymbol{\alpha}$ are linear functions of $\mathbf{r}_i$'s whose means are zero, and conditions (1.-3.) give

$$C^* = n^{1/2}\left[\hat{\boldsymbol{\alpha}}\{\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})\} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi) + \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi) - \boldsymbol{\alpha}\right]$$
$$= n^{1/2}\left\{\frac{\partial \hat{\boldsymbol{\alpha}}(\beta, \phi^*)}{\partial \phi}(\hat{\phi} - \phi) + \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi) - \boldsymbol{\alpha}\right\} = O_p(1). \tag{6.9}$$

Consequently $\sum_{i=1}^n \mathbf{U}_i\{\boldsymbol{\beta}, \boldsymbol{\alpha}^*(\boldsymbol{\beta})\}/n^{1/2}$ is asymptotically equivalent to A^* whose asymptotic distribution is multivariate Gaussian with zero mean and covariate matrix $\lim_{n\to\infty} \mathbf{J}_2/n$ (see Theorem 5.2.1). Finally, it is easy to see that $\sum B_i = o_p(n)$, $C = O_p(1)$ and that $\sum A_i/n$ converges to $-\mathbf{J}_1/n$ as $n \to \infty$. This completes the proof of Theorem 5.2.1.

Now we want to prove Theorem 6.4.2, letting $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_G^T)^T$. If observation i lies in group g, then $i = 1, \ldots, n_g$. If we do not refer to index g, then $i = 1, \ldots, n$. We can apply the same lines as above for Theorem 5.2.1. Now we can re-write B^*

in (6.8) as

$$B^* = \frac{1}{n}\sum_{i=1}^{n} \partial \mathbf{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha} = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g} \partial \mathbf{U}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}_g)/\partial \boldsymbol{\alpha}_g.$$

Now $B^* = o_p(1)$, since $\partial \mathbf{U}_i/\partial \boldsymbol{\alpha}_g$ are linear functions of $\mathbf{r}_i$'s whose means are zero, and conditions 1.-3. of Theorem 6.4.2 give

$$\begin{aligned} C^* &= n^{1/2}\left[\hat{\boldsymbol{\alpha}}\{\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})\} - \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi) + \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi) - \boldsymbol{\alpha}\right] \\ &= n^{1/2}\left\{\frac{\partial \hat{\boldsymbol{\alpha}}(\beta, \phi^*)}{\partial \phi}(\hat{\phi} - \phi) + \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi) - \boldsymbol{\alpha}\right\} \\ &= n^{1/2}\sum_{g=1}^{G} \frac{\partial \hat{\boldsymbol{\alpha}}_g(\beta, \phi^*)}{\partial \phi}(\hat{\phi} - \phi) + (n/n_g)^{1/2}\sum_{g=1}^{G} n_g^{1/2}\left(\hat{\boldsymbol{\alpha}}_g(\boldsymbol{\beta}, \phi) - \boldsymbol{\alpha}_g\right) \\ &= O_p(1) \end{aligned}$$

The remaining lines are the same as above. □

Applicability of Groupwise Correlation Estimation

In the following, we label the groupwise correlation estimation method as *groupwise method* and the method where we assume the same correlations for all observations as *standard method*. What advantages does the groupwise method have? Clearly, we require α_g to be $n_g^{1/2}$ consistent. In other words, we require n_g to be reasonable large. We cannot estimate the correlation structure for each single observation separately, because $n_g = 1$ is certainly not a large number. Often, there are only single observations and the question arises whether, given some grouping, the groupwise method does make sense in terms of better efficiency for the estimation of $\boldsymbol{\beta}$, our primary goal. To answer these questions, we conduct a simulation study in the next section.

6.4.3 Simulation Study

Non-Repeated Multiple Responses

Next, we conduct a simulation study investigating the effect of the chosen working correlation structure and the effect of choosing either the group-wise or the standard (non-groupwise) correlation estimation. We consider the model

$$\text{logit}(\pi_{ij}) = X_{ij}\beta_j,\ i = 1, ..., G\ , j = 1, ..., J \tag{6.10}$$

with $G = 4$ and $J = 3$.

The correlation structure has the following form ($J = 3$)

$$\mathbf{R}_i = \begin{pmatrix} 1 & R_{i12} & \cdots & R_{i1J} \\ R_{i12} & 1 & \cdots & R_{i2J} \\ \vdots & \vdots & & \vdots \\ R_{i1J} & R_{i2J} & \cdots & 1 \end{pmatrix} \equiv \mathbf{R}_{i11}.$$

Let index i of $\mathbf{R}_i$ refer to the ith group, which is sensible, because all observations for a given group have the same probability of a positive response (π_{ij}). Table 6.1 shows the correlation structures considered here.

Table 6.1: Correlation structures for model 6.10

index	$vec(\mathbf{R}_i) = (R_{i12}, R_{i13}, R_{i23})$
1	(-0.1, -0.1, -0.1)
2	(0.1, 0.1, 0.1)
3	(0.3, 0.3, 0.3)
4	(0.5, 0.5, 0.5)
5	(0.1, 0.3, 0.5)
6	(0.2, 0.4, 0.6)
7	(0.1, 0.2, 0.3)
8	(0.3, 0.4, 0.5)

For simplicity, we only consider the three structures: Exchangeable, unstructured and independence.

The odds ratio $\theta_{xy|ik}$ defined by (2.13) on page 60 is another measure of association. From the odds ratio $\theta_{xy|ik}$ and the marginal probabilities $\pi_{x|ik}$ and $\pi_{y|ik}$, we computed the pairwise probability $\pi^{11}_{xy|ik}$, which then determined, with the marginal probabilities, the full pairwise distribution for items x and y. Let Y_x denote whether a subject selects item x. Given group i, if a subject selects item x, then $Y_x = 1$; otherwise, $Y_x = 0$. In a similar way, we can compute the pairwise probability $\pi^{11}_{xy|i}$ from the correlation between Y_x and Y_y and the marginal probabilities $\pi_{x|i}$ and $\pi_{y|i}$, where i refers to the ith group. We have $\text{Cov}(Y_x, Y_y) = \Pr(Y_x = 1, Y_y = 1) - \Pr(Y_x = 1)\Pr(Y_y = 1) = \pi^{11}_{xy|i} - \pi_{x|i}\pi_{y|i}$. By using the formula $\text{Corr}(Y_x, Y_y) = \text{Cov}(Y_x, Y_y)/(\text{Var}(Y_x)^{1/2}\text{Var}(Y_y)^{1/2})$, we can compute $\pi^{11}_{xy|i}$. Then we can compute the other pairwise probabilities $\pi^{01}_{xy|i}$, $\pi^{10}_{xy|i}$ and $\pi^{00}_{xy|i}$ from $\pi^{11}_{xy|i}$, $\pi_{x|i}$ and $\pi_{y|i}$. Finally, we compute the joint probabilities $\tau_{ij'}$ from the complete pairwise distributions for all pairs of items by using the IPF algorithm, as described in Section 2.5 on 59.

We draw $n = 50$ and $n = 200$ observations $\mathbf{y}_i$ randomly from either of the $G = 4$ groups and according to the joint probabilities τ_{ij}, but we require $n_g > 5$ to achieve better convergence, considering that the groupwise method is not applicable for small groupsizes n_g. The covariates X_{ij} were drawn from $N(0, 1)$, but fixed in advance for all simulations, otherwise it would take too long to generate a new joint distribution for all simulated data sets. Then we fit model (6.10) by GEE twice, once using the standard method and once the groupwise method with $G = 4$.

Table 6.2 shows the simulation results for the GEE method and for $\boldsymbol{\beta} = (0.1, 0.2, 0.3)^T$. The first column shows n, the total number of observations generated. The

second column shows the correlation structure for each of the 4 groups. The ith number refers to the ith group's correlation structure $\mathbf{R}_i$, which can be found in Table 6.1 under the index which equals the ith number. For example, if the second number is 4, then the second group has an exchangeable correlation structure with $\alpha = 0.5$, because this is the structure that has index 4 in Table 6.1.

The next columns show the relative efficiency $RE(\hat{\boldsymbol{\beta}})$ for correlation structures unstructured (denoted by "unstr"), exchangeable ("exch") and independence ("ind").

We define the relative efficiency $RE(\hat{\boldsymbol{\beta}})$ of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_J)^T)$ as

$$RE(\boldsymbol{\beta}) = \frac{\sum_{j=1}^{J} \mathbb{E}(\hat{\beta}_j^{TRUE} - \beta_j)^2}{\sum_{j=1}^{J} \mathbb{E}(\hat{\beta}_j - \beta_j)^2} = \frac{\sum_{j=1}^{J} m.s.e.(\hat{\beta}_j^{TRUE})}{\sum_{j=1}^{J} m.s.e.(\hat{\beta}_j)},$$

where $\hat{\beta}_j$ refers to the estimate of β_j for the given working correlation structure and $\hat{\beta}_j^{TRUE}$ stands for the estimated β_j using the correct (true) correlation structure. We use the correct correlation of the simulated distribution and NOT the correct working correlation to estimate the correlation. This ensures that $\hat{\beta}_j^{TRUE}$ has the smallest mean square error. Also, the advantage of our definition is that any other method, such as GEE1 or GEE2, can be easily compared with our method, since relative efficiency of 1.00 is the highest value.

The groupwise and standard methods can also be regarded as part of the working correlation itself, because both methods assume a certain underlying correlation model. The relative efficiency of the method for which the working correlation (the structure and method groupwise/standard) was correctly chosen is denoted by "$*$" in Table 6.1.

However, for some configurations, such as configuration "$1, 1, 4, 4$" for the second column, neither the standard ($G = 1$) nor the groupwise correlation esti-

mation ($G = 4$) is correct, because we simulate two different structures, one for two of the four groups. The latter would be correct for $G = 2$. In this instance, when neither method is correct, we denote the working correlations that are closest to the simulated one by "$^+$".

We simulated $10,000$ data sets for all configurations. The number x in subscript of $RE(\hat{\boldsymbol{\beta}})_x$ is the number of simulations which did not converge for the particular working correlation. The first column also shows the number N for which GEE did not converge for all working correlations including the true correlation. The relative efficiency was computed over $10,000 - N$ data sets only, e.g. line 1 in Table 6.2 shows 50^{348}, meaning that the relative efficiency was computed over $10,000 - 348 = 9,652$ data-sets. GEE did not converge 203 times for the working correlation unstructured (unstr) and 175 times for exchangeable (exch) using the groupwise method, for all other working correlations it did converge for all $10,000$ data sets.

Repeated Multiple Responses

For repeated multiple responses, we consider the model

$$\text{logit}(\pi_{ijt}) = X_{ijt}\beta_j,\ i = 1, ..., G\ , j = 1, ..., J\ , t = 1, ..., T \tag{6.11}$$

Table 6.2: Relative efficiency ($RE(\hat{\boldsymbol{\beta}})$) for model (6.10) and $10,000$ simulated data sets with $\boldsymbol{\beta} = (0.1, 0.2, 0.3)^T$ for non-repeated multiple response data using correlation structures independence (ind), unstructured (unstr) and exchangeable (exch), and the standard and groupwise method ($G = 4$), n stands for number of subjects per data set and N are the number of data sets for which GEE did not converge, the number which is shown for $\mathbf{R}_i$ indicates the correlation structure of Table 6.1 which was used for group $i = 1, \ldots, G$

	correlation	working correlation				
	structure	standard method			groupwise method	
n^N	$\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \mathbf{R}_4$	unstr	exch	ind	unstr	exch
50^{506}	$4, 4, 4, 4$	0.958_0	0.983^*_0	0.662_0	0.907_{346}	0.964_{199}
50^{348}	$1, 1, 4, 4$	0.819_0	0.843^+_0	0.759_0	0.884_{203}	0.948^+_{175}
50^{195}	$1, 2, 3, 4$	0.873_0	0.898_0	0.814_0	0.863_{133}	0.937^*_{77}
50^{246}	$5, 5, 5, 5$	0.959^*_0	0.895_0	0.776_0	0.833_{221}	0.844_{37}
50^{306}	$5, 5, 6, 6$	0.948^+_0	0.868_0	0.693_0	0.834^+_{266}	0.820_{60}
50^{212}	$5, 6, 7, 8$	0.946_0	0.906_0	0.762_0	0.870^*_{174}	0.896_{44}
200^0	$4, 4, 4, 4$	0.993_0	0.999^*_0	0.708_0	0.975_0	0.995_0
200^0	$1, 1, 4, 4$	0.870_0	0.874^+_0	0.784_0	0.973_0	0.991^+_0
200^0	$1, 2, 3, 4$	0.912_0	0.918_0	0.826_0	0.966_0	0.989^*_0
200^0	$5, 5, 5, 5$	0.992^*_0	0.930_0	0.823_0	0.971_0	0.922_0
200^0	$5, 5, 6, 6$	0.986^+_0	0.915_0	0.765_0	0.967^+_0	0.912_0
200^0	$5, 6, 7, 8$	0.979_0	0.930_0	0.799_0	0.969^*_0	0.944_0

*: correct working correlation, $^+$: close to correct

with $G = 4$, $J = 2$, $T = 3$. The correlation matrix has the following form

$$\mathbf{R}_i = \begin{pmatrix} 1 & R_{i11,12} & R_{i12,11} & R_{i12,12} & R_{i13,11} & R_{i13,12} \\ & 1 & R_{i12,21} & R_{i12,22} & R_{i13,21} & R_{i13,22} \\ & & 1 & R_{i22,12} & R_{i23,11} & R_{i23,12} \\ & & & 1 & R_{i23,21} & R_{i23,22} \\ & & & & 1 & R_{i33,12} \\ & & & & & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{i11} & \mathbf{R}_{i12} & \mathbf{R}_{i13} \\ \mathbf{R}_{i12} & \mathbf{R}_{i22} & \mathbf{R}_{i23} \\ \mathbf{R}_{i13} & \mathbf{R}_{i23} & \mathbf{R}_{i33} \end{pmatrix}. \tag{6.12}$$

We consider the following working correlations: First, we regard the vectors $\mathbf{y}_i$ of length $J \cdot T$ as standard multiple response data and choose unstructured (unstr), exchangeable (exch) and independence (ind). Then we simply disregard the time-dependence, only choosing an exchangeable working correlation for J items, but regard observations at different time-points as independent. This structure is identical to a common exchangeable structure for $\mathbf{R}_{itt}$ and a common independence structure for $\mathbf{R}_{it_1t_2}$. We denote this working correlation as "exch(c)-ind". Then we consider the conditional structures exchangeable (items) and unstructured (time), and unstructured (items) and exchangeable (time). The first is denoted by "exch(i) - unstr (t)" and the second by "unstr(i) - exch (t)". We define the relative efficiency in the same way as for non-repeated multiple responses.

Table 6.3 shows the correlation structures being used for the simulated data. The top few lines list the indices j_1, j_2 for the items and t_1, t_2 for the occasions of the elements $R_{t_1t_2,j_1j_2}$ of correlation matrix $\mathbf{R}$ defined in equation (6.12). In this way, we can more easily check which value belongs to which correlation parameter. For convenience, the bottom two lines list the equivalent indices i and j of the elements R_{ij} of matrix $\mathbf{R}$ being expressed as $\mathbf{R} = (R_{ij})_{i,j=1}^{J \cdot T=9}$. For

some people this might be easier to read. The structures 1-4 are exchangeable structures with varying values. The next 4 structures are of the type "unstr(i) - exch (t)", where we assume $R_{t_1t_2,12} = R_{t_1t_2,21}$ to be consistent with $R_{tt,12} = R_{tt,21}$. Structures 9 and 10 are of type "exch(i) - unstr (t)", and structures 11-14 assume independence between items at different occasions, which is "exch(c)-ind". The last 4 structures (15-18) present an unstructured structure.

Table 6.4 shows the relative efficiency for a variety of configurations for $\boldsymbol{\beta} = (0.2, 0.3)^T$ with $n = 50$ simulated observations for each of the $10,000$ datasets.

Table 6.3: Correlation structures for model 2

$$(R_{ij})_{i,j=1}^{J\cdot T=9} = \mathbf{R} = (R_{t_1t_2,j_1j_2})_{j_1,j_2=1;t_1,t_2=1}^{J=3;T=3}$$

	indices j_1j_2 and t_1t_2 of parameters $R_{t_1t_2,j_1j_2}$														
j_1j_2	12,	11,	12,	11,	12,	21,	22,	21,	22,	12,	11,	12,	21,	22,	12
t_1t_2	11,	12,	12,	13,	13,	12,	12,	13,	13,	22,	23,	23,	23,	23,	33
index	parameters $R_{j_1j_2,t_1t_2}$ respectively R_{ij}														
1	-0.1,	-0.1,	-0.1,	-0.1,	-0.1,	-0.1,	-0.1,	-0.1,	-0.1,	-0.1,	-0.1,	-0.1,	-0.1,	-0.1,	-0.1
2	0.1,	0.1,	0.1,	0.1,	0.1,	0.1,	0.1,	0.1,	0.1,	0.1,	0.1,	0.1,	0.1,	0.1,	0.1
3	0.3,	0.3,	0.3,	0.3,	0.3,	0.3,	0.3,	0.3,	0.3,	0.3,	0.3,	0.3,	0.3,	0.3,	0.3
4	0.5,	0.5,	0.5,	0.5,	0.5,	0.5,	0.5,	0.5,	0.5,	0.5,	0.5,	0.5,	0.5,	0.5,	0.5
5	0.1,	0.2,	0.1,	0.2,	0.1,	0.1,	0.3,	0.1,	0.3,	0.1,	0.2,	0.1,	0.1,	0.3,	0.1
6	0.15,	-0.1,	0.15,	-0.1,	0.15,	0.15,	-0.2,	0.15,	-0.2,	0.15,	-0.1,	0.15,	0.15,	-0.2,	0.15
7	-0.1,	0.2,	-0.1,	0.2,	-0.1,	-0.1,	0.1,	-0.1,	0.1,	-0.1,	0.2,	-0.1,	-0.1,	0.1,	-0.1
8	0.1,	-0.1,	0.1,	-0.1,	0.1,	0.1,	-0.2,	0.1,	-0.2,	0.1,	-0.1,	0.1,	0.1,	-0.2,	0.1
9	0.1,	0.2,	0.2,	0.5,	0.5,	0.2,	0.2,	0.5,	0.5,	0.3,	0.4,	0.4,	0.4,	0.4,	0.6
10	0.3,	-0.2,	-0.2,	0.5,	0.5,	-0.2,	-0.2,	0.5,	0.5,	0.4,	-0.1,	-0.1,	-0.1,	-0.1,	0.5
11	0.3,	0,	0,	0,	0,	0,	0,	0,	0,	0.3,	0,	0,	0,	0,	0.3
12	0.5,	0,	0,	0,	0,	0,	0,	0,	0,	0.5,	0,	0,	0,	0,	0.5
13	0.1,	0,	0,	0,	0,	0,	0,	0,	0,	0.1,	0,	0,	0,	0,	0.1
14	0.4,	0,	0,	0,	0,	0,	0,	0,	0,	0.4,	0,	0,	0,	0,	0.4
15	0.00,	0.05,	0.10,	0.15,	0.20,	0.25,	0.30,	0.35,	0.40,	0.45,	0.50,	0.55,	0.60,	0.65,	0.70
16	0.70,	0.65,	0.60,	0.55,	0.50,	0.45,	0.40,	0.35,	0.30,	0.25,	0.20,	0.15,	0.10,	0.05,	0.00
17	0.00,	0.05,	0.10,	0.15,	0.20,	0.25,	0.30,	0.35,	0.70,	0.65,	0.60,	0.55,	0.50,	0.45,	0.40
18	0.70,	0.65,	0.60,	0.55,	0.50,	0.45,	0.40,	0.00,	0.05,	0.10,	0.15,	0.20,	0.25,	0.30,	0.35
ij	12,	13,	14,	15,	16,	23,	24,	25,	26,	34,	35,	36,	45,	46,	56
	indices ij of R_{ij}														

Table 6.4: Relative efficiency ($RE(\hat{\boldsymbol{\beta}})$) for model (6.11) and $10,000$ simulated data sets with $\boldsymbol{\beta} = (0.2, 0.3)^T$ for repeated multiple response data ($T = 3$) using correlation structures independence (ind), unstructured (unstr) and exchangeable (exch), and the standard and groupwise method ($G = 4$), n stands for number of subjects per data set and N are the number of data sets for which GEE did not converge, the number which is shown for $\mathbf{R}_i$ indicates the correlation structure of Table 6.3 which was used for group $i = 1, \ldots, G$

	correlation structure	working correlation										
		standard method						groupwise method				
n	$\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \mathbf{R}_4$	unstr	exch	ind	exch(c)-ind	exch(i)-unstr(t)	unstr(i)-exch(t)	unstr	exch	exch(c)-ind	exch(i)-unstr(t)	unstr(i)-exch(t)
50^{3394}	$1,1,1,1$	0.922_{0}	0.994^{*}_{0}	0.960_{0}	0.954_{0}	0.964_{0}	0.983_{2}	0.710_{1240}	0.978_{22}	0.938_{0}	0.692_{1770}	0.797_{1208}
50^{3675}	$1,2,3,4$	0.835_{0}	0.894_{0}	0.901_{0}	0.878_{0}	0.868_{0}	0.890_{0}	0.657_{1525}	0.995^{*}_{6}	0.882_{20}	0.620_{2627}	0.953_{140}
50^{3115}	$6,6,6,6$	0.933_{0}	0.880_{0}	0.877_{0}	0.869_{0}	0.849_{0}	0.990^{*}_{0}	0.599_{1751}	0.872_{1}	0.847_{0}	0.663_{1372}	0.849_{675}
50^{2710}	$5,6,7,8$	0.864_{0}	0.923_{0}	0.917_{0}	0.914_{0}	0.899_{0}	0.921_{0}	0.626_{1650}	0.924_{1}	0.896_{0}	0.723_{1238}	0.933^{*}_{334}
50^{4982}	$9,9,9,9$	0.934_{2}	0.825_{0}	0.663_{0}	0.701_{0}	0.976^{*}_{0}	0.818_{0}	0.566_{2138}	0.837_{1}	0.688_{13}	0.517_{3615}	0.797_{79}
50^{4763}	$9,9,10,10$	0.900_{0}	0.745_{0}	0.639_{0}	0.711_{0}	0.942^{+}_{0}	0.747_{0}	0.537_{2160}	0.786_{1}	0.686_{8}	0.521^{+}_{3425}	0.747_{86}
50^{3005}	$11,11,11,11$	0.929_{0}	0.917_{0}	0.912_{0}	0.994^{*}_{0}	0.974_{0}	0.917_{0}	0.675_{1232}	0.909_{2}	0.981_{3}	0.672_{2125}	0.878_{80}
50^{3544}	$11,12,13,14$	0.888_{0}	0.832_{0}	0.827_{0}	0.960_{0}	0.941_{0}	0.834_{0}	0.611_{1343}	0.824_{1}	0.982^{*}_{18}	0.605_{2676}	0.790_{92}
50^{5327}	$15,15,15,15$	0.946^{*}_{0}	0.821_{0}	0.742_{0}	0.746_{0}	0.959_{3}	0.828_{0}	0.690_{1730}	0.824_{1}	0.726_{16}	0.426_{4397}	0.807_{79}
50^{5324}	$15,15,16,16$	0.532^{+}_{0}	0.582_{0}	0.583_{0}	0.536_{0}	0.563_{0}	0.580^{+}_{0}	0.562_{1671}	0.573_{1}	0.515_{5}	0.350_{4358}	0.598_{95}
50^{4955}	$15,16,17,18$	0.732_{0}	0.803_{0}	0.672_{0}	0.692_{0}	0.764_{0}	0.796^{*}_{0}	0.821_{1636}	0.800_{0}	0.674_{10}	0.425_{4002}	0.814_{52}

*: correct working correlation, $^{+}$: close to correct

Results

The results confirm previous simulations studies, for instance Liang and Zeger (1986), choosing the correct working correlation gives most efficient parameter estimates. The groupwise method gives more efficient parameter estimates provided different groups have different structures. Similarly, the standard method also yields more efficient parameter estimates provided the correlation is indeed equal for all observations. This was expected, because assuming one correlation structure for either all observations or just for observations within a group also specifies a working correlation. Only when the number of parameters is quite large and the number of observations is quite small, the advantage of choosing, correctly, the groupwise method vanishes. For example, in Table 6.2 for configuration $5, 6, 7, 8$ of the second column, we see that the groupwise method works worse than the standard method, although we expect the opposite. There are two explanations. Either the unstructured working correlation simply has too many parameters or the unstructured working correlation estimated by (5.16) on page 154 does not have exactly the same form as formula (5.12) on page 153 to estimate correlation parameters. Formula (5.16) uses divisor n, whereas according to formula (5.12), it should be $n - p$. The simulations are very computationally expensive. For $n = 50$ and non-repeated multiple response data, one configuration (one line in Table 6.2) - simulating $10,000$ datasets and fitting all of the various GEE methods - requires half a day. Each of the other configurations in Tables 6.2 and 6.4 takes roughly about 5-7 days with standard modern computers available to us.

Generally, when grouped observations with large group sample sizes are given, then we suggest the groupwise method, because the large group sample sizes guarantee good correlation estimates. The more parameters the working correla-

tion requires, the bigger the group sizes have to be to gain efficiency advantages from the groupwise method over the standard method. Otherwise, when we do not assume a subject specific correlation or group sizes are small, the standard method is recommended. Only when a subject specific correlation model is assumed, GEE1 introduced by Prentice (1988) is preferred. Furthermore, only if the correlation model can be trusted or the association/correlation parameters are of primary interest, GEE2 (Zhao and Prentice 1990) is recommended instead.

6.4.4 Missing Data

So far, we have not discussed missing data. Clearly, the STAT 291 data contains missing data. Let the observed responses be denoted by $\mathbf{y}^O$ and the unobserved or missing responses by $\mathbf{y}^M$. If the missingness is independent of both $\mathbf{y}^O$ and $\mathbf{y}^M$, the mechanism is called *missing completely at random* (MCAR). A subcase of MCAR is covariate-dependent missingness (Hedeker and Gibbons 2006), which allows missingness to depend on the observed covariates, e.g. increasing in time. Covariate dependence can also be considered as conditional independence: Given the covariates, the missingness is independent of both $\mathbf{y}^O$ and $\mathbf{y}^M$. Another missingness is termed *missing at random* (MAR), which allows missingness not only to depend on fully observed covariates $\mathbf{x}_i$, but also on the observed responses $\mathbf{y}^O$. In other words, given $\mathbf{x}_i$ and $\mathbf{y}^O$, the missingness is independent of $\mathbf{y}^M$. GEE can only handle data being missing completely at random (MCAR). For the STAT 291 data, covariate dependence (MCAR) seems a reasonable assumption, as it can be ruled out that missingness depends on the students' favourite bar and its features, but will rather depend on covariates as time, the student's major, age, etc. Hence, GEE is applicable and leads to consistent estimates. However, under the weaker assumption of MAR, GEE does not provide consistency

anymore in contrast to ML methods, as introduced previously, and generalised linear mixed models which are introduced later. In the next subsection, a small modification of GEE which can handle MAR is considered.

6.4.5 Weighted Generalised Estimation Equations

Fitzmaurice, Molenberghs and Lipsitz (1995) and Ali and Talukder (2005) considered missing data mechanisms for longitudinal binary data deriving weighted generalised estimation equations (WGEE). Let $D_i = t$ denote the dropout time for given observation i, which is the occasion t from where all data is missing. $T + 1$ represents complete data. The authors modified the score equations (6.4) having the form $\sum_{i=1}^{n} \mathbf{U}_i$ to

$$\sum_{i=1}^{n} \frac{1}{v_{it}} \mathbf{U}_i, \tag{6.13}$$

with weights $\frac{1}{v_{it}}$ where $v_{it} = \Pr(D_i = t|\mathbf{y}_i, \mathbf{X}_i, \gamma)$ is the probability of a dropout of the ith subject on the tth occasion and where γ is some parameter modelling the dropout times. For details we refer to the articles mentioned above. Repeated multiple responses can be considered as multivariate longitudinal binary data, hence, these weighted score equations (6.13) are a useful alternative if MCAR can be ruled out. For the STAT 291 data we do not need WGEE, because MCAR can be assumed; however, for other repeated multiple response data, where MCAR seems an unrealistic assumption but MAR seems sensible, WGEE provides a useful alternative.

6.5 Generalised Linear Mixed Models

The fixed parameters in ordinary GLM or GEE describing the factors effect are independent of the sample. In contrast, generalised linear mixed models (GLMM) additionally include a cluster specific effect, the random effect. As explained in the introduction, the modelling without random effects is called population-averaged modelling, whereas the modelling approach containing the random effects is referred to as subject-specific modelling. Let $\mathbf{u}_i$ be the random effect vector for cluster/observation i and let $\mathbf{Q}_i$ be the design matrix for ith the random effect. Conditional on $\mathbf{u}_i$, the distribution of y_{ijt} is assumed to be from the exponential family type with density $f(y_{ijt}|\mathbf{u}_i;\boldsymbol{\beta})$ and conditional mean $\mu_{ijt} = \mathbb{E}(y_{ijt}|\mathbf{u}_i)$, in our case the distribution is binary and $\mu_{ijt} \equiv \pi_{ijt}$. The linear predictor for a GLMM is

$$g(\pi_{ijt}) = \mathbf{z}_{ijt}^T\boldsymbol{\beta}_{jt} + \mathbf{q}_{ijt}^T\mathbf{u}_i = \eta_{ijt} \tag{6.14}$$

or in vector form

$$\mathbf{g}(\boldsymbol{\pi}_i) = \mathbf{Z}_i\boldsymbol{\beta} + \mathbf{Q}_i\mathbf{u}_i = \boldsymbol{\eta}_i$$

where $\mathbf{Z}_i$, $\mathbf{g}$ and $\boldsymbol{\beta}$ have the same meaning as in model (6.1) and where $\mathbf{Q}_i = (\mathbf{q}_{i11}^T, \ldots, \mathbf{q}_{iJT}^T)^T$. The random effects $\mathbf{u}_i$ of dimension r are assumed to be normal $N(\mathbf{0}, \boldsymbol{\Sigma})$ with unknown positive definite covariance matrix $\boldsymbol{\Sigma}$, where the density is denoted by $f(\mathbf{u}_i;\boldsymbol{\Sigma})$. By the conditional independence, the conditional density of $\mathbf{y}$ given $\mathbf{u}$ has the form

$$f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta}) \text{ with } f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta}) = \prod_{j=1}^{J}\prod_{t=1}^{T} f(y_{ijt}|\mathbf{u}_i;\boldsymbol{\beta}). \tag{6.15}$$

We can also write

$$f(\mathbf{u};\boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(\mathbf{u}_i;\boldsymbol{\Sigma}), \tag{6.16}$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$ and $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)^T$. Note that unconditionally $y_{ij_1t_1}$ and $y_{ij_2t_2}$ are positively correlated (Agresti 2002, p.497).

Now, the likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y})$

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma}) d\mathbf{u} \tag{6.17}$$

is maximised to obtain ML parameter estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. This likelihood function is often called *marginal likelihood* after integrating out the random effects (Agresti 2002).

Maximising the (marginal) likelihood is a ML-method, hence the missing data mechanism allows MAR, in contrast to GEE which only allows the stronger assumption of MCAR. The integral usually cannot be solved analytically and numerical methods must be used. Several approaches maximising (6.17) are discussed next.

6.5.1 Gauss-Hermite Quadrature Methods

Let the random effect $\mathbf{u}_i$ be parameterised by $\mathbf{u}_i = \boldsymbol{\Sigma}^{1/2}\mathbf{a}_i$, with $\boldsymbol{\Sigma}^{1/2}$ being the left Cholesky factor $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})^T$, such that $\mathbf{a}_i$ has mean zero and covariance matrix $\mathbf{I}$, the density is denoted by $\tilde{f}(\mathbf{a}_i)$, and the linear predictor has the form $g(\mu_{it}) = \mathbf{z}_{ijt}^T\boldsymbol{\beta} + \mathbf{q}_{ijt}^T\boldsymbol{\Sigma}^{1/2}\mathbf{a}_i$. Now the likelihood (6.17) does not depend on $\boldsymbol{\Sigma}$, but on the parameter vector $vec(\boldsymbol{\Sigma}^{1/2}) =: \overrightarrow{\boldsymbol{\Sigma}^{1/2}}$ containing the elements of the lower triangular matrix $\boldsymbol{\Sigma}^{1/2}$, which is denoted by $l(\boldsymbol{\alpha}; \mathbf{y})$ with $\boldsymbol{\alpha}^T = (\boldsymbol{\beta}^T, (\overrightarrow{\boldsymbol{\Sigma}^{1/2}})^T)$. When integrating (6.17) with respect to $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)^T$, the reparameterised likelihood

$l(\boldsymbol{\alpha}; \mathbf{y})$ has the form

$$\int_{\mathbb{R}} \dots \int_{\mathbb{R}} \exp(-x_1^2) \cdot \dots \cdot \exp(-x_r^2) v(x) dx_1 \dots dx_r. \tag{6.18}$$

For this type of integral the *Gauss-Hermite* approximation can be applied and the integral can be approximated by

$$l_i(\boldsymbol{\alpha}; \mathbf{y}_i) = \int f(\mathbf{y}_i|\mathbf{a}_i; \boldsymbol{\alpha}) \tilde{f}(\mathbf{a}_i) \mathbf{d}\mathbf{a}_i \approx \sum_{j=1}^{m} w_j f(\mathbf{y}_i|\mathbf{d}_j; \boldsymbol{\alpha}), \tag{6.19}$$

where $\mathbf{d}_j$ is one of the m quadrature points and w_j is the weight associated with $\mathbf{d}_j$. The multivariate case follows from applying the Gauss-Hermite approximation for each dimension separately ($r = 1$) and applying the Cartesian product. For one dimension the quadrature points and the weights follow from the Hermite polynomial. The approximated likelihood or log-likelihood can now be maximised by standard methods, such as Newton-Raphson, to obtain ML estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$. The number of quadrature points m must be large enough to yield accurate ML estimates and this number increases exponentially with dimension r, which becomes infeasible for quite small dimensions. Liu and Pierce (1994) considered adapted Gauss-Hermite quadrature to reduce the number of quadrature points.

6.5.2 Monte Carlo Methods

The simplest *Monte-Carlo* (MC) approximation has the form

$$l_i(\boldsymbol{\alpha}; \mathbf{y}_i) \approx \frac{1}{m} \sum_{j=1}^{m} f(\mathbf{y}_i|\mathbf{d}_{ij}; \boldsymbol{\alpha}) \tag{6.20}$$

where the m values $\mathbf{d}_{ij}$ are drawn from $f(\mathbf{u};\boldsymbol{\Sigma})$. Suppose we can generate samples $\mathbf{d}_{ij}$ from another distribution, the *importance sampling distribution*, with density $h(\cdot)$. Then the following MC approximation, called *importance sampling* (Shao 1999), can also be used

$$l_i(\boldsymbol{\alpha};\mathbf{y}_i) \approx \frac{1}{m}\sum_{j=1}^{m}\frac{f(\mathbf{y}_i|\mathbf{d}_{ij};\boldsymbol{\alpha})f(\mathbf{d}_{ij};\boldsymbol{\Sigma})}{h(\mathbf{d}_{ij})} = \sum_{j=1}^{m} w_{ij} f(\mathbf{y}_i|\mathbf{d}_{ij};\boldsymbol{\alpha}) \tag{6.21}$$

with weights $w_{ij} = \frac{1}{m}\frac{f(\mathbf{d}_{ij};\boldsymbol{\Sigma})}{h(\mathbf{d}_{ij})}$, which might be advantageous if sampling from $f(\mathbf{u};\boldsymbol{\Sigma})$ is difficult.

6.5.3 Estimation of Random Effects

The subject-specific random effects cannot be estimated by applying the ML principle. Applying Bayes' theorem we have

$$f(\mathbf{u}|\mathbf{y};\boldsymbol{\beta},\boldsymbol{\Sigma}) = \frac{f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})f(\mathbf{u};\boldsymbol{\Sigma})}{\int f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})f(\mathbf{u};\boldsymbol{\Sigma})d\mathbf{u}} \propto f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})f(\mathbf{u};\boldsymbol{\Sigma}). \tag{6.22}$$

The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are not known, but replacing them with some consistent estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$, enables us to apply the empirical Bayes' principle (Fahrmeir and Tutz 2001). The "best" Bayesian point estimator (in square error) is the posterior mean $\mathbb{E}(\mathbf{u}|\mathbf{y})$. Also the covariance $\mathrm{Cov}(\mathbf{u}|\mathbf{y})$ is obtainable given the posterior density. For both quantities, generally, integrals must be computed numerically with e.g. Gauss-Hermite or MC.

6.5.4 Indirect Maximisation with EM algorithm

As before, let $\mathbf{y}$ be the observed data and $\mathbf{u}$ be the random effects, which can be considered as unobserved data and let $\boldsymbol{\Psi} = (\boldsymbol{\beta}^T, \overrightarrow{\boldsymbol{\Sigma}}^T)^T$ denote both the model

parameter $\boldsymbol{\beta}$ and the parameter of the covariance $\boldsymbol{\Sigma}$. Assume both $\mathbf{y}$ and $\mathbf{u}$ are observed, the complete likelihood can be expressed as $f(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma})$, hence, the complete log-likelihood is (McCulloch 1997)

$$\log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\beta}) + \log f(\mathbf{u}_i; \boldsymbol{\Sigma}). \tag{6.23}$$

The expectation-maximisation (EM) algorithm has two steps. First let us define

$$Q^{(0)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}') = \mathbb{E}(\log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\Psi})|\mathbf{y}; \boldsymbol{\Psi}') = \int \log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\Psi}) f(\mathbf{u}|\mathbf{y}; \boldsymbol{\Psi}') \mathbf{du}. \tag{6.24}$$

Note that $f(\mathbf{u}|\mathbf{y}; \boldsymbol{\Psi}')$ depends on both $\boldsymbol{\beta}'$ and $\boldsymbol{\Sigma}'$ with $\boldsymbol{\Psi}' = (\boldsymbol{\beta}', \boldsymbol{\Sigma}')$, see (6.22). $\boldsymbol{\Psi}'$ can be seen as an old estimate in an iteration scheme and $\boldsymbol{\Psi}$ as the new estimate. At first the expectation in $Q^{(0)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}')$ (E-step) is computed and at a second step this expression is maximised (M-step) with respect to $\boldsymbol{\Psi}$ for given $\boldsymbol{\Psi}'$. The first term of the complete log-likelihood in (6.23) depends on $\boldsymbol{\beta}$ and the second on $\boldsymbol{\Sigma}$ yielding

$$Q^{(0)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}') = \mathbb{E}(\log f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta})|\mathbf{y}; \boldsymbol{\Psi}') + \mathbb{E}(\log f(\mathbf{u}; \boldsymbol{\Sigma})|\mathbf{y}; \boldsymbol{\Psi}'). \tag{6.25}$$

Therefore the M-step and E-step can be performed separately for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. A dispersion parameter ϕ could also be included in $f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}) = f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}, \phi)$.
Generally, the integral in (6.24) respectively (6.25) must be computed numerically. To approximate the integral numerically by MC approximation, we need to sample from $f(\mathbf{u}|\mathbf{y}; \boldsymbol{\Psi}')$, which can be achieved by the Metropolis-Hasting (MH) algorithm. Let now $h(\cdot)$ denote a candidate distribution and let $\mathbf{u}_i^{k-1}$ be a previous draw from $f(\mathbf{u}_i|\mathbf{y}_i; \boldsymbol{\Psi}')$. Draw a new candidate $\mathbf{u}_i^*$ from $h(\mathbf{u}_i)$. Now accept $\mathbf{u}_i^*$ as the new draw from $f(\mathbf{u}_i|\mathbf{y}_i; \boldsymbol{\Psi}')$ with probability $A_k(\mathbf{u}_i^{k-1}, \mathbf{u}_i^*)$ by setting $\mathbf{u}_i^k := \mathbf{u}_i^*$

with

$$A_k(\mathbf{u}_i^{k-1}, \mathbf{u}_i^*) = \min\left\{1, \frac{f(\mathbf{u}_i^*|\mathbf{y}_i; \mathbf{\Psi}')h(\mathbf{u}_i^{k-1})}{f(\mathbf{u}_i^{k-1}|\mathbf{y}_i; \mathbf{\Psi}')h(\mathbf{u}_i^*)}\right\}.$$

Otherwise reject $\mathbf{u}_i^*$ and accept the existing point instead. Then continue with $k := k + 1$. This procedure still depends on the unknown density $f(\mathbf{u}|\mathbf{y}; \mathbf{\Psi}')$. By setting $h(\mathbf{u}_i) := f(\mathbf{u}_i; \mathbf{\Sigma}')$ the term $A_k(\mathbf{u}_i^{k-1}, \mathbf{u}_i^*)$ simplifies to

$$A_k(\mathbf{u}_i^{k-1}, \mathbf{u}_i^*) = \min\left\{1, \frac{f(\mathbf{y}_i|\mathbf{u}_i^*; \boldsymbol{\beta}')}{f(\mathbf{y}_i|\mathbf{u}_i^{k-1}; \boldsymbol{\beta}')}\right\}$$

and now only depends on the known distribution $f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\beta})$, the conditional likelihood function. Thus, the term $Q^{(0)}(\mathbf{\Psi}|\mathbf{\Psi}')$ can be approximated by sampling a large number m of samples $\mathbf{u}_i^1, \ldots, \mathbf{u}_i^m$ from $f(\mathbf{u}_i|\mathbf{y}_i; \mathbf{\Psi}')$ as described above for all $i = 1, \ldots, n$.

At the second step new estimates for $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$ can be obtained by maximising $Q^{(0)}(\mathbf{\Psi}|\mathbf{\Psi}')$, or equivalently maximising $\mathbb{E}(\log f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta})|\mathbf{y}; \mathbf{\Psi}')$ and $\mathbb{E}(\log f(\mathbf{u}; \mathbf{\Sigma})|\mathbf{y}; \mathbf{\Psi}')$ according to equation (6.25). The maximisation of $\mathbb{E}(\log f(\mathbf{u}; \mathbf{\Sigma})|\mathbf{y}; \mathbf{\Psi}')$ is equivalent to finding the ML estimator for $\mathbf{\Sigma}$ on the "sample" $\mathbf{u}_i^1, \ldots, \mathbf{u}_i^m$. In our case $f(\mathbf{u}; \mathbf{\Sigma})$ is assumed to be multivariate normal and therefore the ML estimator $\hat{\mathbf{\Sigma}}_{ML}$ has a closed form. Also, $f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta})$ is assumed to belong to the exponential family and the ML estimator for $\boldsymbol{\beta}$ can be obtained similarly to the ML estimation of generalised linear models via a Newton-Raphson or Scoring iteration scheme.

McCulloch (1997) proposed several algorithms. One of those, termed Monte-Carlo-Newton-Raphson (MCNR), is:

1. choose starting values $\boldsymbol{\beta}^{(\mathbf{0})}$,$\mathbf{\Sigma}^{(0)}$ ($\phi^{(0)}$), set k:=0

2. generate m values $\mathbf{u}_i^0, \mathbf{u}_i^1, \ldots, \mathbf{u}_i^m$ from the condition distribution $f(\mathbf{u}_i|\mathbf{y}_i; \boldsymbol{\beta}^{(\boldsymbol{k})}, \mathbf{\Sigma}^{(k)}, \phi^{(k)})$ for $i = 1, \ldots, n$ using the MH algorithm

3. calculate $\boldsymbol{\beta^{(k+1)}}$

$$\boldsymbol{\beta^{(k+1)}} = \boldsymbol{\beta^{(k)}} + \hat{\mathbb{E}}[\mathbf{X}^T\mathbf{W}\mathbf{X}|\mathbf{y}]^{-1}\mathbf{X}^T\left(\hat{\mathbb{E}}\left[\mathbf{W}\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}\mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}}(\mathbf{y}-\boldsymbol{\mu})|\mathbf{y}\right]\right)$$

with $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$

4. (optional if ϕ is unknown)
calculate $\phi^{(k+1)}$ that solves $\mathbb{E}(\partial \log f(\mathbf{y}|\mathbf{u};\boldsymbol{\theta})/\partial\phi|\mathbf{y}) = 0$ or with scoring algorithm

5. also determine $\boldsymbol{\Sigma}^{(k+1)}$ which maximises $1/m\sum_{i=1}^{m}\log f(\mathbf{u}^{(k)}|\boldsymbol{\Sigma})$
($\mathbf{u}_i \backsim N(0,\Sigma)$, $\hat{\boldsymbol{\Sigma}}^{k+1} = \frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{n}\mathbf{u}_i^l\mathbf{u}_i^{lT}$)

6. set k:=k+1, if algorithm converged, then proceed with next step (7.), otherwise go back to step 2.

7. consider $\boldsymbol{\beta^{(k+1)}}$, $\phi^{(k+1)}$ and $\boldsymbol{\Sigma}^{(k+1)}$ as ML estimates

The algorithm uses the following notations: $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}, \mathbf{u}) = \mathbb{E}[\mathbf{Y}_i|\mathbf{u}]$, $\mathbf{W}(\boldsymbol{\theta}, \mathbf{u})^{-1}$ $=$ Diag $\left\{(\frac{\partial\boldsymbol{\eta}_i}{\partial\boldsymbol{\mu}_i})^2\text{Var}(\mathbf{Y}_i|\mathbf{u})\right\}$,$\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}$ $=$ Diag$(\frac{\partial\boldsymbol{\eta}_i}{\partial\boldsymbol{\mu}_i})$ and where $\hat{\mathbb{E}}$ denotes the MC approximation of the expectation. Another algorithm without the use of Newton-Raphson, but with maximising $\hat{\mathbb{E}}(\log f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})|\mathbf{y})$ was called MCEM (McCulloch 1997). Both MCEM and MCNR work on the log of the complete likelihood. In contrast, importance sampling (see (6.20) and (6.21)) samples directly from $f(\mathbf{y}_i|\mathbf{u}_i)$ and the likelihood is maximised directly, referred to as simulated maximum likelihood (SML) by McCulloch (1997). McCulloch showed that MCNR and MCEM reach the neighbourhood of the true parameters reasonable fast, but final convergence is achieved slowly. SML only works reasonable well for good starting values and if an optimal importance sampling distribution is used. A hybrid method beginning with MCNR to find good starting values and finishing

with SML by approximating the optimal importance sampling distribution from the given estimates was suggested. Neither method necessarily converges to the global maximum.

Sampling from $f(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\Psi}') \propto f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta}')f(\mathbf{u}_i;\boldsymbol{\Sigma}')$ can also be achieved by *rejection sampling* as suggested by Booth and Hobert (1999) to yield real independent samples. Sample a candidate $\mathbf{u}_i^*$ from $f(\mathbf{u}_i;\boldsymbol{\Sigma}')$ and, independently, another w from the uniform distribution $U[0,1]$. $\mathbf{u}_i^*$ is accepted, if $w \leq f(\mathbf{y}_i|\mathbf{u}_i^*;\boldsymbol{\beta}')/\tau$ with $\tau = \sup_u \{f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta}')\}$. The computation of τ for every iteration is not difficult, which can be seen as a likelihood of a GLM and, hence, is quite easily obtainable. For certain models, τ must only be computed once for the given data. In the case of a very low acceptance rate, the authors also suggest using importance sampling with the multivariate student t-density whose mean and variance match the mode and curvature of $f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta}')f(\mathbf{u}_i;\boldsymbol{\Sigma}')$. It should be the mode and curvature of $f(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\Psi}')$, but this density is unknown. However, $f(\mathbf{u}_i|\mathbf{y}_i;\boldsymbol{\Psi}')$ is proportional to $f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta}')f(\mathbf{u}_i;\boldsymbol{\Sigma}')$ (see (6.22)), such that the maximisation of $f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta}')f(\mathbf{u};\boldsymbol{\Sigma}')$ with the EM-algorithm still yields correct results. The mode and curvature can also be approximated by Lagrange approximations, see Booth and Hobert (1998) and Booth and Hobert (1999). Booth and Hobert (1999) also considered the MC error which influences the MC approximation of the integral, which depends on the choice of m. Let us define

$$Q^{(1)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}') = \frac{\partial}{\partial\boldsymbol{\Psi}}Q^{(0)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}'), Q^{(2)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}') = \frac{\partial^2}{\partial\boldsymbol{\Psi}\partial\boldsymbol{\Psi}^T}Q^{(0)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}') \tag{6.26}$$

and

$$S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi}) = \frac{\partial}{\partial\boldsymbol{\Psi}}\log f(\mathbf{y},\mathbf{u};\boldsymbol{\Psi}).$$

They showed $\boldsymbol{\Psi}$ is approximately normal distributed with mean $\boldsymbol{\Psi}^*$ and covari-

ance

$$\text{Cov}(\boldsymbol{\Psi}|\boldsymbol{\Psi}') \approx Q^{(2)}(\boldsymbol{\Psi}^*|\boldsymbol{\Psi}')^{-1}\mathbb{E}(S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi})S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi})^T|\mathbf{y};\boldsymbol{\Psi}')Q^{(2)}(\boldsymbol{\Psi}^*|\boldsymbol{\Psi}')^{-1}, \quad (6.27)$$

where $\boldsymbol{\Psi}^*$ satisfies $Q^{(1)}(\boldsymbol{\Psi}^*|\boldsymbol{\Psi}') = 0$. Booth and Hobert (1999) suggest constructing an approximate $100(1-\alpha)$ confidence region for $\boldsymbol{\Psi}^*$ after the $(r+1)$th iteration using $\text{Cov}(\boldsymbol{\Psi}^{(r+1)}|\boldsymbol{\Psi}^{(r)})$ in (6.27). If the previous value $\boldsymbol{\Psi}^{(r)}$ lies in this region, the authors suggest increasing m to $m := m + m/3$ with $\alpha = 0.25$. Another advantage of their algorithm is that the information matrix is a by-product. Louis (1982) showed

$$I(\boldsymbol{\Psi}|\mathbf{y}) := -\frac{\partial^2 l}{\partial\boldsymbol{\Psi}\partial\boldsymbol{\Psi}^T} = -Q^{(2)}(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}) - \text{Cov}(S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi})|\mathbf{y};\hat{\boldsymbol{\Psi}}) \quad (6.28)$$

evaluated at $\hat{\boldsymbol{\Psi}}$. At $\hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{\Psi}}'$ we have

$$\text{Cov}(S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi})|\mathbf{y};\hat{\boldsymbol{\Psi}})|_{\boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}} = \mathbb{E}(S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi})S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi})^T|\mathbf{y};\hat{\boldsymbol{\Psi}})|_{\boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}}$$

because $\mathbb{E}(S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})|_{\boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}} = 0$. We can also write

$$I(\boldsymbol{\Psi}|\mathbf{y}) = \text{Cov}(S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi});\hat{\boldsymbol{\Psi}}) - \text{Cov}(S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi})|\mathbf{y};\hat{\boldsymbol{\Psi}}),$$

which is simply the difference between the unconditional and conditional variance. The quantities $Q^{(2)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(r)})$ and $S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi})$ have the following form

$$\begin{aligned} Q^{(2)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}^{(r)}) &= \begin{pmatrix} \mathbb{E}(\frac{\partial^2}{\partial\boldsymbol{\beta}\boldsymbol{\beta}^T}\log f(\mathbf{y},\mathbf{u};\boldsymbol{\beta}|\mathbf{y};\boldsymbol{\Psi}^{(r)})) & \mathbf{0} \\ \mathbf{0} & \mathbb{E}(\frac{\partial^2}{\partial\overrightarrow{\boldsymbol{\Sigma}}\overrightarrow{\boldsymbol{\Sigma}}^T}\log f(\mathbf{u};\boldsymbol{\Sigma}|\mathbf{y};\boldsymbol{\Psi}^{(r)})) \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{E}(\mathbf{X}^T\mathbf{W}\mathbf{X}|\mathbf{y};\boldsymbol{\Psi}^{(r)}) & \mathbf{0} \\ \mathbf{0} & -\frac{n}{2}(\boldsymbol{\Sigma}^{(r+1)})^{(-1)}\otimes(\boldsymbol{\Sigma}^{(r+1)})^{(-1)} \end{pmatrix} \end{aligned}$$

and

$$S(\mathbf{y},\mathbf{u};\boldsymbol{\Psi}) = \begin{pmatrix} \frac{\partial}{\partial\boldsymbol{\beta}}\log f(\mathbf{y},\mathbf{u};\boldsymbol{\beta}) \\ \frac{\partial}{\partial\overrightarrow{\boldsymbol{\Sigma}}}\log f(\mathbf{u};\boldsymbol{\Sigma}) \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{W}(\beta,\mathbf{u})\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}\mid_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(k)}}(\mathbf{y}-\boldsymbol{\mu}) \\ -\frac{n}{2}\boldsymbol{\Sigma}^{(-1)} + \frac{1}{2}\boldsymbol{\Sigma}^{(-1)}\sum_{i=1}^{n}\mathbf{u_i}\mathbf{u_i}^T\boldsymbol{\Sigma}^{(-1)} \end{pmatrix}.$$

The dispersion parameter ϕ was omitted for simplification. For details of the EM algorithm, see Little and Rubin (1987).

Parameter Transformation

The parameters transformation from $\boldsymbol{\beta}$ and $\overrightarrow{\boldsymbol{\Sigma}}$ to $\boldsymbol{\alpha}^T = (\boldsymbol{\beta}^T, (\overrightarrow{\boldsymbol{\Sigma}}^{1/2})^T)$ as described above has several advantages. In (6.19) the random effect density does not depend on any parameters, because it is simply multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_r$. The linear predictor can be written as

$$\eta_{it} = \left[\mathbf{q}_{ijt}^T\mathbf{a}_i^T \otimes \mathbf{z}_{ijt}^T\right]\begin{bmatrix} \boldsymbol{\beta} \\ \overrightarrow{\boldsymbol{\Sigma}}^{1/2} \end{bmatrix}.$$

The parameter vector $\boldsymbol{\alpha}$ consists of all unknown fixed parameters and is included in the conditional likelihood $f(\mathbf{y}_i|\mathbf{a}_i;\boldsymbol{\alpha})$ such that the iteration scheme might look slightly easier because of the absence of estimating the fixed parameters $\boldsymbol{\Sigma}$ of the random effect density $f(\mathbf{u};\boldsymbol{\Sigma})$. For details see e.g. Tutz and Hennevogl (1996) and Fahrmeir and Tutz (2001, Chapter 7).

6.5.5 Approximate Likelihood Methods

Approximate maximum likelihood methods are based on first- and second order Taylor series expansions of the likelihood. Marginal Quasi-likelihood (MQL) involves expansion around the fixed part of the model, whereas penalised quasi-

likelihood (PQL) also includes the random part in its expansion. For example, Stiratelli et al. (1984), Schall (1991) and Breslow and Clayton (1993) derived the following *penalised log-likelihood* equations

$$l(\boldsymbol{\beta}, \mathbf{u}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\mathbf{u}_i, \boldsymbol{\beta}) - 1/2 \sum_{i=1}^{n} \mathbf{u}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{u}_i, \tag{6.29}$$

although Stiratelli et al. (1984) derived these equations by using a Bayesian approach and Schall (1991) by using the BLUP procedure. MQL (Zeger et al. 1988, Goldstein 1991) focuses on the marginal relationship between covariates and outcomes.

However, all these approaches can yield poor estimates, which can be severely biased, in particular for first order expansions (Breslow and Lin 1995). More recently, Raudenbush et al. (2000) introduced a fast method combining a fully multivariate Taylor series expansion and a Laplace approximation, yielding accurate results.

6.5.6 Bayesian Mixed Models

For the Bayesian approach, the prior distributions for all parameters $f(\boldsymbol{\beta}|\cdot)$, $f(\mathbf{u}|\cdot)$ and $f(\boldsymbol{\Sigma}|\cdot)$ must be specified. The posterior distribution is $f(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Sigma}|\mathbf{y})$. Sampling from the posterior distribution enables us to obtain parameter estimates, which can be accomplished by applying e.g. Gibbs sampling (Fahrmeir and Tutz 2001). Gibbs sampling is an easy iterative scheme to sample from $f(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Sigma}|\mathbf{y})$. First, set any two starting value $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$ and $\mathbf{u}^{(0)}$, without loss of generality, let the first two be given. Set $k := 0$. Now sample $\mathbf{u}^{(k)}$ from $f(\mathbf{u}|\boldsymbol{\beta}^{(k)}, \boldsymbol{\Sigma}^{(k)})$. Then sample $\boldsymbol{\beta}^{(k+1)}$ from $f(\boldsymbol{\beta}|\mathbf{u}^{(k)}, \boldsymbol{\Sigma}^{(k)})$ and $\boldsymbol{\Sigma}^{(k+1)}$ from $f(\boldsymbol{\Sigma}|\mathbf{u}^{(k)}, \boldsymbol{\beta}^{(k+1)})$ and set $k := k + 1$. Stop with $k = N$.The triple $(\mathbf{u}^{(k)}, \boldsymbol{\Sigma}^{(k)}, \boldsymbol{\beta}^{(k)}), k = 1, \ldots, N$ represent a sample of

size N from the posterior $f(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\Sigma}|\mathbf{y})$. The (posterior) means from this sample are regarded as the estimates, e.g. $\hat{\boldsymbol{\beta}} = 1/N \sum_{k=1}^{N} \boldsymbol{\beta}^{(k)}$.

6.5.7 Semi- or Nonparametric Maximum Likelihood EM algorithm

Instead of assuming that the distribution of the random effects $f(\mathbf{u})$ follows any parametric distribution such as the multivariate normal distribution, one can also assume that $f(\mathbf{u})$ is any discrete distribution with probabilities $\mathbf{p} = (p_1, \ldots, p_K)$ with finite support size K and mass points $\mathbf{m} = (m_1, \ldots, m_K)$. If no assumptions can be made, $\mathbf{p}, \mathbf{m}$ and K are unknown. When applying the EM algorithm, it may happen that for fixed K some mass points equal $\pm\infty$, which corresponds to cell probabilities ± 1. Hartzel, Agresti and Caffo (2001) discuss this approach and suggest successively fitting while increasing K. There is no big difference in estimates between the parametric EM algorithm and the non-parametric approach. They suggest using the non-parametric approach to check whether the estimates from the parametric and non-parametric EM algorithm are approximate equal, otherwise this could be a sign for the inadequacy of the model.

6.6 Stat 291 Data

The Stat 291 data has $T = 3$ time-points and has various bar features to be considered as items. For simplicity, we consider $J = 3$ items only, namely "drink deals" (item 1), "pool table" (item 2) and "sports TV" (item 3). The repeated multiple responses were created by assigning a positive response at occasion t for item j (e.g. "drink deals"), when the student's favourite bar at occasion t has a certain feature (e.g. "drink deals") to be also considered as item j. We refer

to the introduction (Section 6.1 on page 184), where we explained in detail how to obtain the repeated multiple responses from the subject's most favourite bar and its features. We use the logit link, because the marginal responses are binary, and consider a common effect, such that marginal model (6.1) has form (6.2). We tested several models by excluding/including step by step those variables that are highly insignificant and those variables whose exclusion makes other variables insignificant. Finally we ended up with a quite simple model with variables "work" (question 13: working=1/ not working=0), "friends" (5a: yes=1/ no=0) and "sex" (male=1/ female=0) for item 1, "pool" (4: yes=1/ no=0) and "sex" for item 2 and finally variable "smoke" (12: yes=1/ no=0) and "sex" for item 3.

We use the same covariates for the random effects model, which is of the form

$$\log\left(\frac{\pi_{ijt}}{1-\pi_{ijt}}\right) = \alpha_j + \mathbf{x}_{it}^T\boldsymbol{\beta}_j + \mathbf{x}_{i0}^T\boldsymbol{\beta}_{0j} + u_{ij}$$

with $\mathbf{u}_i = (u_{i1}, \ldots, u_{iJ})^T$. Random effect vector $\mathbf{u}_i \in \mathbb{R}^J$ is assumed to be multivariate normal, referring to subject i, where the jth component u_{ij} refers to the jth item. In the literature, as in Agresti and Liu (2001), often only one single univariate random effect is used to account for dependency between items, but this seems too stringent. On the contrary, allowing JT correlated random effects, one for each component, does not seem appropriate either. The random effect structure was chosen in such a way, because we expect that the π_{it} vary more over items than over time.

Table 6.5 shows the parameter estimates for various GEE and GLMM methods. For GEE, a good structure seems "unstr(i)-exch(t)", because the exact structure between items is very unclear and seems rather heterogeneous between all pairs of items (unstructured), whereas for the time-dependence, we can assume

one of the typical time-structures, such as exchangeable. Unfortunately, GEE only converges for the standard method. Another structure introduced earlier and considered in the simulation study is "exch(i)-unstr(t)", for which GEE converges also for the groupwise (G) method. For the groupwise method, we use the $G = 2$ groups formed by variable "sex", which is an explanatory variable for the marginal responses of all 3 items. It seems sensible that "sex" is an also an explanatory variable for the correlation. For the groupwise method, variable "pool" has a smaller p-value than for the standard method. For instance, for the structure "exch(i)-unstr(t)", the standard method yields the p-value 0.168 (not significant) and the groupwise method gives a p-value of 0.052 (marginally significant). Without applying the groupwise method, variable "pool" would remain undetected.

For the random effect model, we applied the MCNR algorithm (McCulloch 1997) in combination with confidence regions (Booth and Hobert 1999). The parameter estimates are very similar to the estimates of GEE using structure independence. GLMM can only impose non-negative correlations between items, but it is very unlikely that all of the $1/2(J \cdot T) \times (J \cdot T - 1)$ correlation parameter of the multiple responses are non-negative, in fact for an unstructured correlation structure, GEE gives correlation parameter estimates ranging from -0.121 to $+0.254$ indicating relatively small correlations, both positive and negative. Therefore, the method adjusts the random effects to be small imposing only small non-negative correlations, which is then close to the independence structure. We also fitted the model with penalised quasi likelihood (PQL). Parameter estimates for PQL differ from the other estimates, indicating a bias (as mentioned earlier in Subsection (6.5.5) for quasi-likelihood methods) and providing unreliable estimates.

Let us discuss the parameter estimates for structure "exch(i)-unstr(t)" and $G = 2$. The odds of selecting a bar offering drink deals (Pool Table/ Sports

TV) are $1/\exp(-0.645) = 1/0.52 = 1.91$ ($1.58/\ 2.31$) times higher for females than for males. Females seem to be more aware of the bar's features and select a bar as most favourite based on the bar's features. The odds for working people choosing a bar that offers drink deals are $\exp(0.553) = 1.77$ ($\exp(0.575) = 1.78$) times those for non-working people (for people who go out to socialise than for those who don't). Also the odds of selecting a bar offering a pool table are $\exp(-0.792) = 0.45$ times for those who enjoy playing pool than for those who don't. We probably would expect the opposite, but eventually the pool table is not of high importance for selecting a most favourite bar for those who do enjoy playing pool. The method PQL also suggest variable pool to be marginally significant, but we regard PQL generally as unreliable. We must consider the possibility that variable pool is simply insignificant.

For people who smoke the odds of selecting a bar offering some sorts of Sports TV are $\exp(0.381) = 1.46$ times those for people not smoking. This is not too unexpected, because some people might see a link between Sports TV and smoking.

6.7 Discussion

In this chapter, we mainly focused on GEE and GLMM methods for modelling repeated multiple responses due to the impractical nature of the ML approach. Although both methods seem similar and contain the same fixed effect parameters $\boldsymbol{\beta}$, they are not identical unless $\boldsymbol{\Sigma} = \mathbf{0}$ for GLMM and an independence structure is chosen for GEE. ML estimation does not need any assumption about correlation parameters, however, the method becomes infeasible for small J and T due to the 2^{JT} joint probabilities. In addition, it requires non-zero joint cell counts, an almost impractical condition to be met for $JT \geq 6$, because of the sparseness

Table 6.5: Parameter estimates (s.e.) and p-value for GEE and GLMM models

	Drink Deals			Pool Table		Sports TV	
method	work	friends	sex	pool	sex	smoke	sex
GEE	0.488	0.013	-0.641	-0.843	-0.473	0.201	-0.609
unstr(i)-ex(t)	(0.286)	(0.512)	(0.329)	(0.498)	(0.355)	(0.212)	(0.385)
	0.087	0.980	0.052	0.091	0.182	0.342	0.114
GEE ($G = 2$)	0.553	0.575	-0.645	-0.792	-0.460	0.381	-0.837
ex(i)-unstr(t)	(0.241)	(0.295)	(0.323)	(0.408)	(0.329)	(0.188)	(0.393)
	0.022	0.051	0.046	0.052	0.162	0.042	0.033
GEE	0.439	0.480	-0.685	-0.519	-0.523	0.396	-0.674
ex(i)-unstr(t)	(0.215)	(0.323)	(0.324)	(0.377)	(0.335)	(0.196)	(0.401)
	0.041	0.136	0.034	0.168	0.118	0.043	0.093
GEE	0.540	0.655	-0.766	-0.207	-0.478	0.298	-0.599
ind	(0.279)	(0.497)	(0.269)	(0.370)	(0.278)	(0.291)	(0.340)
	0.053	0.187	0.004	0.575	0.085	0.306	0.079
GEE	0.436	0.479	-0.673	-0.223	-0.498	0.262	-0.635
unstr	(0.250)	(0.421)	(0.311)	(0.336)	(0.346)	(0.208)	(0.380)
	0.081	0.255	0.030	0.507	0.149	0.206	0.095
GEE	0.556	0.528	-0.746	-0.322	-0.549	0.250	-0.706
ex ($G = 2$)	(0.269)	(0.402)	(0.323)	(0.364)	(0.355)	(0.219)	(0.385)
	0.039	0.189	0.021	0.375	0.122	0.252	0.067
GEE	0.541	0.528	-0.759	-0.294	-0.545	0.248	-0.692
ex	(0.268)	(0.407)	(0.321)	(0.361)	(0.354)	(0.220)	(0.388)
	0.043	0.195	0.018	0.414	0.124	0.260	0.075
GLMM	0.544	0.662	-0.775	-0.209	-0.487	0.299	-0.609
MCNR	(0.280)	(0.498)	(0.270)	(0.371)	(0.278)	(0.291)	(0.341)
mult	0.051	0.183	0.004	0.571	0.080	0.305	0.074
GLMM	0.796	0.527	-0.996	-0.592	-0.926	0.253	-1.134
PQL	(0.228)	(0.409)	(0.659)	(0.321)	(0.661)	(0.227)	(0.686)
uni	0.001	0.198	0.131	0.065	0.161	0.265	0.099

of the data. If zero cell counts occur, the estimates are called "extended ML estimates".

In this chapter, we did not consider log-linear models as another ML method. Loglinear models often provide a better fit than the random effect models, because they do not impose severe restrictions on the joint distribution of $\mathbf{y}_i$. However, they cannot describe within-subject effects, in contrast to random effect models. Also, they cannot describe how the probabilities of a positive response depend on the covariates, which is the basic concept of our modelling approach. Interpretation of log-linear models is another difficulty. In addition, log-linear models do specify the cell counts of the joint tables and share the same limitation as the ML method; both deal with $2^{JT} - 1$ parameters per joint table. Due to these difficulties, we do not consider log-linear models as useful for modelling of repeated multiple response data.

Unconditionally, GLMM does impose a correlation structure on the components of $\mathbf{y}_i$. However, these correlation parameters are non-negative. The larger the diagonal elements of $\boldsymbol{\Sigma}$ are, the larger are these non-negative correlations. This imposed model assumption might be severely violated for a given data set due to negative correlations between observations and might lead to too small estimates for Diag($\boldsymbol{\Sigma}$), an indication of model misspecification. If this occurs, the parameter estimates $\hat{\boldsymbol{\beta}}$ might be inaccurate.

In our view, GEE is the preferable method. It is widely implemented in all common statistical packages and a simple choice of the correlation structure as exchangeable yields more efficient estimates than the GLM approach assuming independence between all items. If one wishes to obtain even more efficient estimates, we recommend implementing the GEE procedure with some of the considered more sophisticated correlation structures, for instance autoregressive

(time) and unstructured (items). If the data is grouped, then we also recommend estimating the structure for different groups separately provided groupsizes are reasonable large (> 10).

Regarding missing data (Subsection 6.4.4 on page 212), Little (1995) noted that covariate-dependence should stand alone and not be thought as a special case of MCAR, because usually MCAR stands for the missingness being not only independent of the observed and unobserved responses, but of the whole data including the covariates (Little and Rubin 1987). Hence, there might be a little confusion about the meaning of MCAR. GEE works for the assumption of covariate-dependence, hence, it is in our view the preferred method in regards to missing data.

Chapter 7

Graphical Model-Checking Techniques for the Proportional Odds Model

7.1 Introduction

Ordered categorical variables occur in many applications. In this chapter, we consider two examples. Table 7.1 shows the data given by Neter, Wasserman and Kutner (1985, Chapter 9), where an agronomist studied the effects of moisture (X_1, in inches) and temperature (X_2, in ^0C) on the yield of a new hybrid tomato (Y), which is divided into three levels: high (1) , medium (2), and low (3). Particularly in the health sciences, ordinal scales are very common. Often there are clinical reasons for recording certain continuous measurements in an ordinal scale. One such example is the Normative Aging Study (NAS), where 682 men aged of 48 to 93 years reported their medical examination, such as fasting blood glucose (FBG) and two markers of systemic inflammation, namely, white blood cell count

(*wbc*) and blood levels of C-reactive protein (*crp*). FBG is is often recorded in three categories, clinically defined as "normal level" (level 1), "impaired level" (level 2), and "diabetic level"(level 3).

Table 7.1: The yield of a new hybrid tomato

observation	1	2	3	4	5	6	7	8	9
X_1 (Moist)	6	6	6	6	6	8	8	8	8
X_2 (Temp)	20	21	22	23	24	20	21	22	23
Y	1	2	2	1	2	1	1	1	1
observation	10	11	12	13	14	15	16	17	18
X_1 (Moist)	8	10	10	10	10	10	12	12	12
X_2 (Temp)	24	20	21	22	23	24	20	21	22
Y	2	1	1	1	2	2	2	2	2
observation	19	20	21	22	23	24	25		
X_1 (Moist)	12	12	14	14	14	14	14		
X_2 (Temp)	23	24	20	21	22	23	24		
Y	2	2	3	3	3	3	3		

Effects of treatment or any other covariates like age, ethnicity on such ordinal responses can be studied through the multivariate GLM methodology. Let Y be J-category ordinal response variable and $\mathbf{x}$ be a column vector of linear predictors. The proportional odds model

$$\text{logit}[P(Y \leq j \mid \mathbf{x})] = \alpha_j - \mathbf{x}^T\boldsymbol{\gamma},\ j = 1, ..., J-1, \tag{7.1}$$

which uses logits of cumulative probabilities, is currently the most popular model. Model (7.1) implies that the cumulative odds ratio referring to two sets of linear predictors is constant for all categories j. It also does not depend on the scores assigned, a major advantage when compared to other existing ordinal models (Agresti 2002, Chapter 7), so different studies assigning different scores still yield similar conclusions.

Testing the adequacy of the proportional odds model can be done in several ways. As we pointed out in the introduction (Section 1.4 on page 32), many methods fit the partial proportional odds model

$$\text{logit}[P(Y \leq j \mid \mathbf{x})] = \alpha_j - \mathbf{x}^T \boldsymbol{\gamma}_j, \; j = 1, ..., J-1,$$

and test whether the $c-1$ effect parameters $\boldsymbol{\gamma}_j$ are equal, for example Peterson and Harrell (1990) and Brant (1990) propose Wald and score tests for testing $\boldsymbol{\gamma}_1 = \cdots = \boldsymbol{\gamma}_{J-1}$. Another method was proposed by Lipsitz et al. (1996), who generalised the popular Hosmer–Lemeshow statistic, originally introduced by Hosmer and Lemeshow (2000) for checking the adequacy of a logistic regression model, to the situation of ordinal response models. Toledano and Gatsonis (1996) applied a receiver operating characteristic (ROC) curve which plots sensitivity against 1 - specificity for all possible collapsings of the J categories.

All of the above methods check the overall adequacy of the proportional odds model. They do not give a close view of model mis-specification for the functional form of specific covariates.

In standard linear regression models, plotting residuals versus an explanatory variable X is often viewed as a diagnostic tool to examine model mis-specification in X. The residuals for a binary logistic model are typically defined as the difference between observed response, and the estimated probability of the response, conditional on the covariates. The plot of the residuals versus X is hard to interpret in such cases.

Su and Wei (1991) considered a cumulative residual process assessing only the overall adequacy of a GLM. Lin et al. (2002) extended their idea and presented graphical methods for assessing the adequacy of the functional form of one or

more covariates, the functional form of the linear predictor and the overall adequacy of the model by considering similar cumulative residual processes. Their methods are based on the GEE methodology, which includes GLM and multivariate GLM as special cases, and allow quite simple and easy interpretation of diagnostical plots showing the cumulative residual processes. Recently, Arbogast and Lin (2005) also proposed such cumulative residual processes for case-control studies using logistic regression.

The current chapter generalises their methods of checking model mis-specification in the context of the proportional odds model for $J > 2$ using two different routes.

One approach considers the proportional odds model as $J-1$ logistic regression models, where the response categories are collapsed into the binary outcome $(\leq j, > j)$, $j = 1, \ldots, J-1$. The cumulative sums of residuals have the same form as the ones given by Arbogast and Lin (2005) for each of the collapsed logistic models. In the second approach, the proportional odds model (7.1) is viewed as a member of the class of multivariate GLM, where the response variable is a vector of indicator responses $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i,J-1})^T$, where $y_{ij} = 1$ if subject i falls in category j and is 0 otherwise. Consequently, the residual, the difference between the observed value of the response and the predicted probability of the response for the ith subject, is a $(J-1) \times 1$ vector. We consider a multivariate cumulative residual process consisting of multivariate residuals to assess model mis-specification, that converges to a multivariate Gaussian process. We can also apply the univariate residual processes proposed by Lin et al. (2002) to the vector responses $\mathbf{y}_i$. This process is identical to the sum over the components of our multivariate cumulative residual process.

The remainder of the chapter is organised as follows. In Sections 7.2 and 7.3,

we introduce the two new approaches, the binary and multivariate, respectively. In Section 7.4 we conduct a simulation study investigating the relative performance of the proposed methods. Section 7.5 illustrates the methods on the two examples: (1) The agronomist study (Neter et al. 1985, chapter 9) measuring the yield of a new hybrid tomato with effects moisture and temperature (see Table 7.1), and (2) the recent dataset from the Normative Aging Study (Bell, Rose and Damon 1966) which studies the effect of the white blood cell count (wbc) and the C-reactive protein (crp) on fasting blood glucose (FBG) measurement. The last section finishes with some concluding remarks also discussing the applicability of multiple response data. We published these sections in a very similar form (Liu et al. 2008). However, the article does not contain such detailed proofs and also does not illustrate the methods in the agronomist study. The PhD candidate's work of the paper was to find proofs for the proposed new multivariate methods, to conduct simulation studies to compare methods, to apply these methods to the examples, creating graphics, and write part of the text.

7.2 Binary Approach

Now we consider the first approach considering the proportional odds model as $J-1$ logistic regression models. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i,J-1})^T$ be the response for subject i, where $i = 1, \ldots, n$. If the subject responds as level j, then $y_{ij} = 1$ and $y_{ih} = 0$ for all $h \neq j = 1, \ldots, J-1$. If the response is at baseline level J, then $\mathbf{y}_i = (0, 0, \ldots, 0)^T$.

We first define the collapsed responses as $y^*_{ij} = \sum_{h=1}^{j} y_{ih}$, where $j = 1, \ldots, J-1$. That is, y^*_{ij} is a binary response variable having values 1, or 0. It can be considered as a binary outcome when we collapse the response categories into ($\leq j$,

$> j$), $j = 1, \ldots, J-1$. If the response category is $\leq j$, then $y_{ij}^* = 1$. Otherwise, $y_{ij}^* = 0$. For the jth collapsing, the residual r_{ij}^* is defined as

$$r_{ij}^* = y_{ij}^* - P(Y \leq j \mid \mathbf{x}_i), \tag{7.2}$$

where we assume $\mathbf{x}_i$ is a column vector of predictors for the ith subject and $P(Y \leq j \mid \mathbf{x}_i)$ satisfies the proportional odds model (7.1), which is simply a logistic regression model for a fixed j of the form $\log(\pi_{ij}^*/(1-\pi_{ij}^*)) = \mathbf{z}_{ij}^T \boldsymbol{\beta}_j$ with $\mathbf{z}_{ij} = \left[1\,,\, \mathbf{x}_i^T\right]^T$ and $\boldsymbol{\beta}_j^T = (\alpha_j, \boldsymbol{\gamma}^T)$. Therefore, this approach is equivalent to the method used for the logistic regression model given by Arbogast and Lin (2005) for each specific collapsing. Consider the following stochastic process

$$W_k^{(j)}(t; \hat{\boldsymbol{\beta}}_j) = n^{-1/2} \sum_{i=1}^{n} \hat{r}_{ij}^* \mathbb{1}(x_{ik} \leq t), \tag{7.3}$$

where x_{ik} is the kth component of $\mathbf{x}_i$. $\mathbb{1}(\cdot)$ is the indicator function, which equals one if the expression in brackets is true, otherwise it is zero. The form $W_k^{(j)}(t; \hat{\boldsymbol{\beta}}_j)$ uses a cumulative sum of the residuals $\hat{r}_{ij}^*$ over the values of x_{ik}. Following Arbogast and Lin's argument, under the null hypothesis $\mathcal{H}_0$ that model (7.1) is correct, $W_k^{(j)}(t; \hat{\boldsymbol{\beta}}_j)$ converges weakly to a zero-mean Gaussian process. The distribution of the Gaussian process can be approximated by that of

$$\widehat{W}_k^{(j)}(t; \hat{\boldsymbol{\beta}}_j) = n^{-1/2} \sum_{i=1}^{n} \left\{ \mathbb{1}(x_{ik} \leq t) + \hat{\boldsymbol{\eta}}^T(t, \hat{\boldsymbol{\beta}}_j) \left[n^{-1} \mathcal{I}(\hat{\boldsymbol{\beta}}_j) \right]^{-1} \mathbf{z}_{ij} \right\} N_i \, \hat{r}_{ij}^* \,, \tag{7.4}$$

where

$$\begin{aligned} \hat{\boldsymbol{\eta}}(t, \hat{\boldsymbol{\beta}}_j) &= n^{-1/2} \partial W_k^{(j)}(t; \boldsymbol{\beta}_j) / \partial \boldsymbol{\beta}_j \\ &= -n^{-1} \sum_{i=1}^{n} \hat{P}(Y \leq j \mid \mathbf{x}_i) \left[1 - \hat{P}(Y \leq j \mid \mathbf{x}_i) \right] \mathbb{1}(x_{ik} \leq t) \mathbf{z}_{ij}, \end{aligned}$$

and where $\mathcal{I}(\hat{\boldsymbol{\beta}}_j)$ is the information matrix, and $\{N_i,\, i = 1, \ldots, n\}$ are independent standard normal random variables. The proof of this result was given in Arbogast and Lin (2005).

Now we can plot the observed cumulative residuals along with a large number of simulated realisations based on the Gaussian process (7.4) and compare their pattern to detect some model mis-specification. Relatively large observed cumulative residuals indicate a violation of the model. Arbogast and Lin (2005) used the Kolmogorov-type supremum statistic $G_{W_k} := \sup_{t\in\mathbb{R}}|W_k(t;\hat{\boldsymbol{\beta}}_j)|$, where $\mathbb{R}$ denotes the real line and W_k stands for $W_k^{(j)}$, $j = 1, \ldots, J-1$ in our case.

Let g_{W_k} denote the observed value of the supremum statistic G_{W_k}. We cannot compute the p-value $\Pr(G_{W_k} \geq g_{W_k})$ of the test directly, but $\Pr(G_{W_k} \geq g_{W_k})$ can be approximated by $\Pr(G_{\widehat{W}_k} \geq g_{W_k})$, where $G_{\widehat{W}_k} = \sup_{t\in\mathbb{R}} |\widehat{W}_k(t;\hat{\boldsymbol{\beta}}_j)|$. Then $\Pr(G_{\widehat{W}_k} \geq g_{W_k})$ is estimated by generating a large number (≥ 1000) of realisations $\widehat{W}_k(t;\hat{\boldsymbol{\beta}}_j)$. That is, the p-value of the test is obtained by computing the proportion of the simulated realisations greater than the largest value of $|W_k^{(j)}(t;\hat{\boldsymbol{\beta}}_j)|$ over t, because the extreme values of $W_k^{(j)}(t;\hat{\boldsymbol{\beta}}_j)$ would suggest that functional mis-specification exists for covariate x_{ik}. Each collapsed response results in a single plot and a single p-value. In total, there are $J-1$ plots denoted by $B_1, \ldots, B_{J-1}$. One might use the Bonferroni method to adjust for the significance level while combining inference from all these plots, so that the overall Type I error rate is less than or equal to the sum of the individual error rates for all $J-1$ plots. The Bonferroni adjusted significance level is thus the significance level divided by $J-1$. Later, we refer to it as Bonf(B).

Our main focus is on the mis-specification of the functional form of a covariate. Arbogast and Lin (2005) also provided the residual processes $W_o^{(j)}$ to assess the overall adequacy of the model and $W_p^{(j)}$ to assess the adequacy of the link

function. The processes have the same form as (7.3) and (7.4) only replacing indicator function $\mathbb{1}(x_{ik} \leq t)$ by $\mathbb{1}(\mathbf{z}_{ij}^T \hat{\boldsymbol{\beta}}_j \leq t)$ for $W_p^{(j)}$ and by $\mathbb{1}(\mathbf{z}_{ij} \leq \mathbf{t})$ for $W_o^{(j)}$, where $\mathbf{z}_{ij} \leq \mathbf{t}$ is true if $z_{ik} \leq t_k$ for all components z_{ijk} of $\mathbf{z}_{ij}$.

7.3 Multivariate Approach

In this section, we propose the multivariate approach based on the multivariate residuals. First, we introduce generalised estimating equations (GEE) and the results for another univariate approach by Lin et al. (2002) formulated in Theorem 7.3.2. Then, we introduce the new multivariate approach and present results in Theorem 7.3.3. In the subsection thereafter, we apply the several proposed processes based on the multivariate approach for the proportional odds model, which can also be expressed as a multivariate generalised linear model (MGLM), a subclass of GEE. Finally, we make some comments about efficient implementation of the processes.

7.3.1 Generalised Equation Equations

Lin et al. (2002) proposed graphical diagnostic methods for generalised estimation equations (GEE), which were introduced by Liang and Zeger (1986). We use the same notations as above, but we use the more general setting that the length of observations may differ. Let y_{ij} be the (not necessarily ordinal) response of the ith subject ($i = 1, \ldots, n$) at the jth occasion ($j = 1, \ldots, J_i$) with $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ_i})^T$. Similarly define the mean $\boldsymbol{\mu}_i$ and the residuals $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\mu}_i$. Let $\mathbf{x}_{ij}$ be the covariates for the jth occasion of ith subject and let $J = \max(J_1, \ldots, J_n)$ be the maximal cluster length. We assume the marginal mean $\mathbb{E}y_{ij} = \mu_{ij}$ depends on the

p-dimensional column vector $\mathbf{z}_{ij}$ by

$$g_j(\mu_{ij}) = \mathbf{z}_{ij}^T \boldsymbol{\beta}_j, \tag{7.5}$$

with unknown parameter vector $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp_j})^T$ of length p_j. Suppose $\mathbf{z}_{ij}$, the ijth contribution to design matrix $\mathbf{Z}$, depends on the covariates $\mathbf{x}_{ij}$. The model can also be expressed in the more compact form

$$\boldsymbol{g}(\boldsymbol{\mu}_i) = \mathbf{Z}_i \boldsymbol{\beta}. \tag{7.6}$$

For more details of GEE, we refer to Section 5.2.2 on page 150.

Remark 7.3.1. Lin et al. (2002) assumed the model $g(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}_j$. However, for GEE the model function g does not need to be identical for all j and may be replaced by g_j. Also, the design matrix may include some entries not being identical to a covariate, but may only depend on them, hence we replace $\mathbf{x}_{ij}$ by $\mathbf{z}_{ij}$ to account for a more general setting.

Empirical Processes

Let us define the empirical processes

$$W_o(\mathbf{t}; \mathbf{b}, \boldsymbol{\beta}) = n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \mathbb{1}(\mathbf{t} - \mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t}) r_{ij}(\boldsymbol{\beta}), \tag{7.7}$$

$$\begin{aligned} &\widehat{W}_o(\mathbf{t}; \mathbf{b}, \boldsymbol{\beta}) \\ &= n^{-1/2} \sum_{i=1}^{n} \left[\sum_{j=1}^{J_i} \mathbb{1}(\mathbf{t} - \mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t}) r_{ij}(\boldsymbol{\beta}) + \boldsymbol{\eta}_{W_o}^T(t, \boldsymbol{\beta}) \boldsymbol{\Omega}(\boldsymbol{\beta})^{-1} \mathbf{U}_i(\boldsymbol{\beta}) \right] N_i \end{aligned} \tag{7.8}$$

with

$$\boldsymbol{\eta}_{W_o}(\mathbf{t};\mathbf{b},\boldsymbol{\beta}) = n^{-1/2}\partial W_o/\partial\boldsymbol{\beta} = -n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\mathbb{1}(\mathbf{t}-\mathbf{b}<\mathbf{z}_{ij}\leq\mathbf{t})\mathbf{M}_{ij}(\boldsymbol{\beta}), \qquad (7.9)$$

where $\mathbf{M}_{ij}(=\partial\mu_{ij}^T/\partial\boldsymbol{\beta})$ is the jth column of $\mathbf{M}_i(=\partial\boldsymbol{\mu}_i^T/\partial\boldsymbol{\beta})$ and $\boldsymbol{\Omega} = n^{-1}\sum_{i=1}^{n}\mathbf{M}_i$ $\mathbf{V}_i^{-1}\mathbf{M}_i^T = n^{-1}\mathbf{J}_1$. The quantities $\mathbf{U}_i$, $\mathbf{M}_i$, $\mathbf{V}_i$ and $\mathbf{J}_1$ were defined in Section 5.2.2 on page 150.

Vector $\mathbf{b}$ is constant and $\mathbb{1}(\mathbf{t}-\mathbf{b}<\mathbf{z}_{ij}\leq\mathbf{t})$ reduces to $\mathbb{1}(\mathbf{z}_{ij}\leq\mathbf{t})$ for $\mathbf{b} = (\infty,\ldots,\infty)$. Define the processes $W_k(t;b,\boldsymbol{\beta}) := W_o(\mathbf{t};\mathbf{b},\boldsymbol{\beta})$ and $\widehat{W}_k(t;b,\boldsymbol{\beta}) := \widehat{W}_o(\mathbf{t};\mathbf{b},\boldsymbol{\beta})$ with $\mathbf{t} = (t_1,\ldots,t_p)$ and $\mathbf{b} = (b_1,\ldots,b_p)$ where $t_l = \infty$ and $t_l - b_l = -\infty$ for $l \neq k$ and $t_l = t$ and $b_l = b$ for $l = k$.

Let the processes $W_p(\mathbf{t};\mathbf{b},\boldsymbol{\beta})$ and $\widehat{W}_p(\mathbf{t};\mathbf{b},\boldsymbol{\beta})$ be similarly defined as $W_o(\mathbf{t};\mathbf{b},\boldsymbol{\beta})$ and $\widehat{W}_o(\mathbf{t};\mathbf{b},\hat{\boldsymbol{\beta}})$ only replacing $\mathbb{1}(\mathbf{t}-\mathbf{b}<\mathbf{z}_{ij}\leq\mathbf{t})$ by $\mathbb{1}(t-b<\mathbf{z}_{ij}\boldsymbol{\beta}_j\leq t)$, where it occurs in the definition of (7.7), (7.8) and (7.9). The Kolmogorov-type supremum statistics are defined as for the binary approach: $G_{W_o} = \sup_{\mathbf{t}\in\mathbb{R}^p}|W_o(\mathbf{t};\mathbf{b},\hat{\boldsymbol{\beta}})|$, $G_{W_p} = \sup_{t\in\mathbb{R}}|W_p(t;\mathbf{b},\hat{\boldsymbol{\beta}})|$ and $G_{W_k} = \sup_{t\in\mathbb{R}}|W_k(t;b,\hat{\boldsymbol{\beta}})|$.

Theorem 7.3.2 (Lin et al. 2002). *Under $\mathcal{H}_0$, that model (7.5) holds, the processes of any of the following pairs*

1. *$W_o(\mathbf{t};\mathbf{b},\hat{\boldsymbol{\beta}})$ and $\widehat{W}_o(\mathbf{t};\mathbf{b},\hat{\boldsymbol{\beta}})$*
2. *$W_p(t;b,\hat{\boldsymbol{\beta}})$ and $\widehat{W}_p(t;b,\hat{\boldsymbol{\beta}})$*
3. *$W_k(t;b,\hat{\boldsymbol{\beta}})$ and $\widehat{W}_k(t;b,\hat{\boldsymbol{\beta}})$,*

are asymptotically equivalent and converge weakly to the same zero-mean Gaussian process. The Kolmogorov-type supremum statistic

(i) G_{W_o} is consistent against any departures from model (7.5)

(ii) G_{W_p} *is consistent against mis-specification of the link function* $\mathbf{g}$

(iii) G_{W_k} *is consistent against mis-specification of the functional form of* z_{ijk}.

Let $\mathcal{H}_1^o$, $\mathcal{H}_1^p$ and $\mathcal{H}_1^k$ denote the alternatives of the null hypothesis $\mathcal{H}_0$, under which the tests G_{W_o}, G_{W_p} and G_{W_k} are consistent against, see (i), (ii) and (iii) of Theorem 7.3.2. The processes W_o, W_p and W_k fluctuate around zero as $\mathbf{t}$ (respectively t) varies. Large values of W (using any subscript p, o or k) indicate a violation of $\mathcal{H}_0$ and that the alternative $\mathcal{H}_1$ might be true. As for the binary approach, we can also plot the observed cumulative residuals W along with a large number of simulated realisations $\widehat{W}$ and see how large the observed cumulative residuals W are relative to the realisations of $\widehat{W}$ and conclude in favour of either $\mathcal{H}_0$ or $\mathcal{H}_1$.

Extension to Multivariate Residuals and Processes

Although process W refers to a multivariate model, the residual process sums over the components of the multivariate residual to obtain a univariate and not multivariate cumulative residual process. In some instances, it might be wiser to consider a multivariate cumulative residual process. Now we extend the univariate processes to such multivariate processes and formulate results in Theorem 7.3.3. Let us define

$$\mathbf{W}_o(\mathbf{t};\mathbf{b},\boldsymbol{\beta}) = n^{-1/2}\sum_{i=1}^{n}\mathbb{I}(\mathbf{t}-\mathbf{b}<\mathbf{Z}_i\leq\mathbf{t})\mathbf{r}_i(\boldsymbol{\beta}) \tag{7.10}$$

$$\widehat{\mathbf{W}}_o(\mathbf{t};\mathbf{b},\boldsymbol{\beta}) = n^{-1/2}\sum_{i=1}^{n}\left[\mathbb{I}(\mathbf{t}-\mathbf{b}<\mathbf{Z}_i\leq\mathbf{t})\mathbf{r}_i(\boldsymbol{\beta}) + \boldsymbol{\eta}^T_{\mathbf{W}_o}(\mathbf{t},\boldsymbol{\beta})\boldsymbol{\Omega}^{-1}(\boldsymbol{\beta})\mathbf{U}_i(\boldsymbol{\beta})\right]N_i \tag{7.11}$$

with

$$\boldsymbol{\eta}_{\mathbf{W}_o}(\mathbf{t};\mathbf{b},\boldsymbol{\beta}) = n^{-1/2}\partial\mathbf{W}_o/\partial\boldsymbol{\beta} = -n^{-1}\sum_{i=1}^{n}\mathbf{M}_i\mathbb{I}(\mathbf{t}-\mathbf{b}<\mathbf{Z}_i\leq\mathbf{t}) \tag{7.12}$$

and $\mathbb{I}(\mathbf{t}-\mathbf{b}<\mathbf{Z}_i\leq\mathbf{t}) := \text{Diag}\{\mathbb{1}(\mathbf{t}-\mathbf{b}<\mathbf{z}_{i1}\leq\mathbf{t}),\ldots,\mathbb{1}(\mathbf{t}-\mathbf{b}<\mathbf{z}_{iJ_i}\leq\mathbf{t})\}$. Define processes $\mathbf{W}_p(\mathbf{t};\mathbf{b},\boldsymbol{\beta})$, $\widehat{\mathbf{W}}_p(\mathbf{t};\mathbf{b},\boldsymbol{\beta})$, $\mathbf{W}_k(t;b,\boldsymbol{\beta})$ and $\widehat{\mathbf{W}}_k(t;b,\boldsymbol{\beta})$ similarly to before with subscripts p and k. Let $G_{\mathbf{W}_o} = \sup_{\mathbf{t}\in\mathbb{R}^p}\|\mathbf{W}_o(\mathbf{t};\mathbf{b},\hat{\boldsymbol{\beta}})\|$ where $\|\cdot\|$ denotes any norm on $\mathbb{R}^J$, similarly $G_{\mathbf{W}_p}$ and $G_{\mathbf{W}_k}$. Such a norm can be seen as a projection to the real plane. Generally, we can consider a continuous function $h(\cdot)$

$$h:\ \mathbb{R}^{J-1}\ \rightarrow\ \mathbb{R}\,,$$

where $\mathbb{R}^{J-1}$ denotes the $(J-1)-$dimensional real plane. Applying function h to the vector of stochastic processes $\mathbf{W}_o$ yields an univariate process.

The following theorem can be seen as an extension of Theorem 7.3.2:

Theorem 7.3.3. *Under $\mathcal{H}_0$, the processes of any of the following pairs*

1. $\mathbf{W}_o(\mathbf{t};\mathbf{b},\hat{\boldsymbol{\beta}})$ *and* $\widehat{\mathbf{W}}_o(\mathbf{t};\mathbf{b},\hat{\boldsymbol{\beta}})$
2. $\mathbf{W}_p(t;b,\hat{\boldsymbol{\beta}})$ *and* $\widehat{\mathbf{W}}_p(t;b,\hat{\boldsymbol{\beta}})$
3. $\mathbf{W}_k(t;b,\hat{\boldsymbol{\beta}})$ *and* $\widehat{\mathbf{W}}_k(t;b,\hat{\boldsymbol{\beta}})$,

are asymptotically equivalent and converge weakly to the same multivariate zero-mean Gaussian process. The tests $G_{\mathbf{W}_o}$, $G_{\mathbf{W}_p}$ and $G_{\mathbf{W}_k}$ are consistent against the same alternatives $\mathcal{H}_1^o$, $\mathcal{H}_1^p$ and $\mathcal{H}_1^k$ (see Theorem 7.3.2). $h(\mathbf{W})$ and $h(\widehat{\mathbf{W}})$ still converge weakly to the same process (not necessarily Gaussian) provided the function h with $h(\mathbf{W})\subset\mathbb{R}$ is almost surely continuous using any of the subscripts o, g and k. If additionally $h(\mathbf{0})=\mathbf{0}$ and from $|\mathbf{c}|<|\mathbf{d}|$ it follows that $|h(\mathbf{c})|<|h(\mathbf{d})|$ (monotonicity condition), the tests

$G_{h(\mathbf{W}_o)}$, $G_{h(\mathbf{W}_p)}$, and $G_{h(\mathbf{W}_k)}$ are still consistent under $\mathcal{H}_1$ against the aforementioned alternatives .

Remark 7.3.4. If from $\mathbf{c} < \mathbf{d}$ it follows that $h(\mathbf{c}) < h(\mathbf{d})$, then h is called strictly monotone or order preserving. For multivariate comparisons "$<$" stands for the product order: $(c_1, \ldots, c_K) = \mathbf{c} < \mathbf{d} = (d_1, \ldots, d_K)$ iff $c_1 < d_1, \ldots, \mathbf{c}_K < \mathbf{d}_K$, similarly $|\mathbf{c}|$ stands for $(|c_1|, \ldots, |c_K|)$.

Unlike the binary approach, we cannot plot the observed multivariate residuals directly, because $\mathbf{W}(t; \hat{\boldsymbol{\beta}})$ is a vector. According to Theorem 7.3.3, the processes $h(\mathbf{W})$ and $h(\widehat{\mathbf{W}})$ are still consistent under the alternative $\mathcal{H}_1$, if function h fulfils the monotonicity condition.

There are several options available for the choice of function $h(\cdot)$. This chapter suggests the following simple choices all fullfiling the monotonicity condition

$$\begin{aligned} \text{sum}(\mathbf{W}) := &\quad h(\mathbf{W}) = \textstyle\sum_{j=1}^{J-1}(W)_j \\ \max(\mathbf{W}) := &\quad h(\mathbf{W}) = \max|\mathbf{W}| \\ \text{prod}(\mathbf{W}) := &\quad h(\mathbf{W}) = \textstyle\prod_{j=1}^{J-1}(W)_j \end{aligned}$$

where $(W)_j$ is the jth component of the vector $\mathbf{W}$.

In addition, the p-value of the test can be calculated in the same way as in the binary approach using a Bonferroni adjustment. We plot the observed multivariate residuals $\mathbf{r}$ with the simulated realisations separated by rows to create $J-1$ plots, denoted by $(\mathbf{W})_1$, ..., $(\mathbf{W})_{J-1}$. If the model is correct, the null hypothesis is accepted for each of the plots. We can adjust the significance level so that the overall Type I error rate is less than or equal to the sum of the individual error rates for all $J-1$ plots. It leads to another diagnostic method denoted by Bonf($\mathbf{W}$).

Proof of Theorem 7.3.3

Before we start with the proof of Theorem 7.3.3, we discuss the applicability of a useful theorem (Theorem 7.3.5) and propose an alternative (Theorem 7.3.8) based on Proposition 7.3.7, which is then used for proving Theorem 7.3.3.

Let D^m be the m-dimensional space of Cadlag functions, right-continuous functions with an existing left limit, and C^m be the m-dimensional space of continuous functions, both defined on $[0, 1]$.

Theorem 7.3.5 (Davidson 1994, p.491)**.** *Let* $\mathbf{W}_n \in D^m$ *be an* m*-vector of random elements.* $\mathbf{W}_n \rightarrow_d \mathbf{W}$*, where* $\mathbb{P}(\mathbf{W} \in C^m) = 1$*, iff* $\boldsymbol{\lambda}^T \mathbf{W}_n \rightarrow_d \boldsymbol{\lambda}^T \mathbf{W}$ *for every fixed* $\boldsymbol{\lambda}$ *with* $\boldsymbol{\lambda}^T \boldsymbol{\lambda} = 1$.

We cannot directly apply this theorem, an extension of the Cramer-Wold theorem for stochastic processes, because not all entries of the design matrix are purely continuous and in $[0, 1]$. However, both are only technical matters, because we can assume without loss of generality that the entries of the design matrix $\mathbf{Z}$ are in $[0, 1]$, and we can partition $\mathbf{Z}$ into $\mathbf{Z}^1$ and $\mathbf{Z}^2$, the discrete and continuous entries of $\mathbf{Z}$. The processes can be re-written as a double sum over all observations and over all possible outcomes generated from $\mathbf{Z}^1$ following a similar approach to Su and Wei (1991). We now cite a proposition and derive a similar theorem to Theorem 7.3.5.

We introduce some new notations that are only used to show the theorem that follows, which is needed to prove Theorem 7.3.3. Let $\{X_{n,i} : i \leq n, n \geq 1\}$ be a triangular array of $\mathcal{W}$-valued random variables, where $\mathcal{W} \subset \mathbb{R}^J$. Also let $\mathcal{T}$ be a pseudometric space with pseudometric ρ. Let $\mathcal{M} = \{\mathbf{f}(\cdot, \tau) : \tau \in \mathcal{T}\}$ be a class of functions ($\in \mathbb{R}^s$) defined on $\mathcal{W}$ and indexed by $\mathcal{T}$. Let us define the empirical process $\mathbf{W}_n = n^{-1/2} \sum_{i=1}^n (\mathbf{f}(X_{n,i}) - \mathbb{E}\mathbf{f}(X_{n,i}))$.

Definition 7.3.6. $\{\mathbf{W}_n, n \geq 1\}$ is stochastically equicontinuous if $\forall \epsilon > 0$ and $\forall \eta > 0$, $\exists \delta > 0$ such that

$$\overline{\lim_{n\to\infty}} \mathbb{P}\left[\sup_{\tau_1,\tau_2\in\mathcal{T}:\rho(\tau_1,\tau_2)<\delta} \|\mathbf{W}_n(\tau_1) - \mathbf{W}_n(\tau_2)\|_2 > \eta\right] < \epsilon \tag{7.13}$$

where $\|\cdot\|_2$ denotes the Euclidian norm.

Proposition 7.3.7 (Andrews 1994, p.2251)**.** *If (i) $(\mathcal{T}, \tau)$ is a totally bounded pseudometric space, (ii) finite dimensional convergence holds: $\forall$ finite subsets $(\tau_1, \ldots, \tau_J)$ of $\mathcal{T}$, $(\mathbf{W}_n(\tau_1)^T, \ldots, \mathbf{W}_n(\tau_J)^T)^T$ converges in distribution, and (iii) $\{\mathbf{W}_n, n \geq 1\}$ is stochastically equicontinuous, then there exists a $\mathcal{B}(\mathcal{T})$-valued (the class of bounded functions on $\mathcal{T}$) stochastic process $\mathbf{W}(\cdot)$ whose sample paths are uniformly ρ continuous with probability one, such that $\mathbf{W}_n(\cdot) \to_d \mathbf{W}(\cdot)$. Conversely, if $\mathbf{W}_n(\cdot) \to_d \mathbf{W}(\cdot)$ and (i) holds, then (ii) and (iii) hold.*

Now we formulate a similar theorem to Theorem 7.3.5:

Theorem 7.3.8. *Let $(\mathcal{T}, \tau)$ be a totally bounded pseudometric space. $\mathbf{W}_n \to_d \mathbf{W}$, iff $\boldsymbol{\lambda}^T\mathbf{W}_n \to_d \boldsymbol{\lambda}^T\mathbf{W}$ for every fixed $\boldsymbol{\lambda}$ with $\boldsymbol{\lambda}^T\boldsymbol{\lambda} = \|\boldsymbol{\lambda}\|_2 = 1$, where $\mathbf{W}$ is a stochastic process whose sample paths are uniformly ρ continuous with probability one.*

Proof of Theorem 7.3.8. "⇒": Let $\mathbf{W}_n \to_d \mathbf{W}$. We apply the continuous mapping theorem to the continuous functional $h(\mathbf{x}) = \boldsymbol{\lambda}^T\mathbf{x}$ and hence $\boldsymbol{\lambda}^T\mathbf{W}_n \to_d \boldsymbol{\lambda}^T\mathbf{W}$. We could also argue as follows:

$$\begin{aligned}
&\|\boldsymbol{\lambda}^T\mathbf{W}_n(\tau_1) - \boldsymbol{\lambda}^T\mathbf{W}_n(\tau_2)\|_2 = |\boldsymbol{\lambda}^T\mathbf{W}_n(\tau_1) - \boldsymbol{\lambda}^T\mathbf{W}_n(\tau_2)| \\
=&|\boldsymbol{\lambda}^T(\mathbf{W}_n(\tau_1) - \mathbf{W}_n(\tau_2))| \leq \|\boldsymbol{\lambda}\|_2\|\mathbf{W}_n(\tau_1) - \mathbf{W}_n(\tau_2)\|_2 \\
=&\|\mathbf{W}_n(\tau_1) - \mathbf{W}_n(\tau_2)\|_2
\end{aligned}$$

and therefore

$$\begin{aligned}&\overline{\lim_{n\to\infty}}\,\mathbb{P}\left[\sup\|\boldsymbol{\lambda}^T\mathbf{W}_n(\tau_1)-\boldsymbol{\lambda}^T\mathbf{W}_n(\tau_2)\|_2>\eta\right]\\ \leq&\overline{\lim_{n\to\infty}}\,\mathbb{P}\left[\sup\|\mathbf{W}_n(\tau_1)-\mathbf{W}_n(\tau_2)\|_2>\eta\right]<\epsilon,\end{aligned}$$

which means $\{\boldsymbol{\lambda}^T\mathbf{W}_n, n\geq 1\}$ is equicontinuous and hence by Proposition 7.3.7: $\boldsymbol{\lambda}^T\mathbf{W}_n\to_d\boldsymbol{\lambda}^T\mathbf{W}$.

"$\Leftarrow$": Now let $\boldsymbol{\lambda}^T\mathbf{W}_n\to_d\boldsymbol{\lambda}^T\mathbf{W}$. By Proposition 7.3.7 $\{\boldsymbol{\lambda}^T\mathbf{W}_n, n\geq 1\}$ is equicontinuous. We want to show $\mathbf{W}_n\to_d\mathbf{W}$ and apply Proposition 7.3.7. From the Cramer-Wold theorem [if for fixed $\boldsymbol{\lambda}$ and τ the random variable $\boldsymbol{\lambda}^T\mathbf{W}(\tau)$ converges in distribution to $\boldsymbol{\lambda}^T\mathbf{W}(\tau)$, then $\mathbf{W}(\tau)$ converges in distribution to $\mathbf{W}(\tau)$] follows (ii). It remains to show (iii), the equicontinuity of $\{\mathbf{W}_n, n\geq 1\}$. Let $\mathbf{e}_i$ be the ith unit vector. For an arbitrary vector $\mathbf{a}$, we can write $\mathbf{a}=\sum_{i=1}^q a_i\mathbf{e}_i$ and $\|a_i\mathbf{e}_i\|_2=\|\mathbf{e}_i^T\mathbf{a}\|_2$. By replacing $\boldsymbol{\lambda}$ by $\mathbf{e}_i$, ϵ by ϵ/J and η by η/J, because $\boldsymbol{\lambda}$, ϵ and η are arbitrary in the definition of equicontinuity, we set $\overline{\lim_{n\to\infty}}\,\mathbb{P}\left[\sup\|\mathbf{e}_i^T\mathbf{W}_n(\tau_1)-\mathbf{e}_i^T\mathbf{W}_n(\tau_2)\|>\eta/J\right]\leq\epsilon/J$. Also let $\mathbf{W}_n=(W_{n,1},\ldots,W_{n,J})^T$.

Now we have

$$\begin{aligned}&\|\mathbf{W}_n(\tau_1)-\mathbf{W}_n(\tau_2)\|_2=\|\sum_{i=1}^J W_{n,i}(\tau_1)\mathbf{e}_i-W_{n,i}(\tau_2)\mathbf{e}_i\|_2\\ \leq&\sum_{i=1}^J\|W_{n,i}(\tau_1)\mathbf{e}_i-W_{n,i}(\tau_2)\mathbf{e}_i\|_2=\sum_{i=1}^J\|(W_{n,i}(\tau_1)-W_{n,i}(\tau_2))\mathbf{e}_i\|_2\\ =&\sum_{i=1}^J\|\mathbf{e}_i^T(\mathbf{W}_n(\tau_1)-\mathbf{W}_n(\tau_2))\|_2=\sum_{i=1}^J\|\mathbf{e}_i^T\mathbf{W}_n(\tau_1)-\mathbf{e}_i^T\mathbf{W}_n(\tau_2)\|_2.\end{aligned}$$

Hence

$$\overline{\lim_{n\to\infty}}\,\mathbb{P}\left[\sup\|\mathbf{W}_n(\tau_1)-\mathbf{W}_n(\tau_2)\|_2>\eta\right]$$

$$
\begin{aligned}
&\leq \overline{\lim_{n\to\infty}} \mathbb{P}\left[\sup \sum_{i=1}^{J} \|\mathbf{e}_i^T \mathbf{W}_n(\tau_1) - \mathbf{e}_i^T \mathbf{W}_n(\tau_2)\|_2 > \eta\right] \\
&\leq \overline{\lim_{n\to\infty}} \mathbb{P}\left[\bigcup_{i=1}^{J} \{\sup \|\mathbf{e}_i^T \mathbf{W}_n(\tau_1) - \mathbf{e}_i^T \mathbf{W}_n(\tau_2)\|_2 > \eta/J\}\right] \\
&\leq \sum_{i=1}^{J} \overline{\lim_{n\to\infty}} \mathbb{P}\left[\sup \|\mathbf{e}_i^T \mathbf{W}_n(\tau_1) - \mathbf{e}_i^T \mathbf{W}_n(\tau_2)\|_2 > \eta/J\right] \\
&< \sum_{i=1}^{J} \epsilon/J = \epsilon.
\end{aligned}
$$

The second inequality follows from $\{\sup \| \sum_{i=1}^{J} a_i\|_2 > \eta\} \subset \bigcup_{i=1}^{J} \{\sup \|a_i\|_2 > \eta/J\}$. Thus, $\{\mathbf{W}_n, n \geq 1\}$ is equicontinuous and from Proposition 7.3.7 follows $\mathbf{W}_n \rightarrow_d \mathbf{W}$. We can also show the equicontinuity more easily. Andrews (1994, p. 2267) noted that equicontinuity for $\{\mathbf{W}_n, n \geq 1\}$ follows from univariate equicontinuity (for $W_{n,i}$). The components of $\{\mathbf{W}_n, n \geq 1\}$ are equicontinuous, because $\{\boldsymbol{\lambda}^T \mathbf{W}_n, n \geq 1\}$ is equicontinuous and by setting $\boldsymbol{\lambda} := \mathbf{e}_j$ the process $\boldsymbol{\lambda}^T \mathbf{W}_n$ equals $(\mathbf{W}_n)_j$. □

Proof of Theorem 7.3.3. Equivalence of $h(\mathbf{W})$ and $h(\widehat{\mathbf{W}})$:

We show that $\mathbf{W}_o(\mathbf{t}; \mathbf{b}, \hat{\boldsymbol{\beta}})$ and $\widehat{\mathbf{W}}_o(\mathbf{t}; \mathbf{b}, \hat{\boldsymbol{\beta}})$ (or more shortly simply $\mathbf{W}_o$ and $\widehat{\mathbf{W}}_o$) converge to a multivariate zero-mean Gaussian process. $\mathbf{W}_k$ and $\widehat{\mathbf{W}}_k$ are special cases of $\mathbf{W}_o$ and $\widehat{\mathbf{W}}_o$ and do not require a separate proof. $\mathbf{W}_p$ and $\widehat{\mathbf{W}}_p$ can be proved similarly only replacing arguments of the indicator functions. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)^T$ be arbitrary but fixed with $\|\boldsymbol{\lambda}\|_2 = 1$ and define $\boldsymbol{\lambda}_i = (\lambda_1, \ldots, \lambda_{J_i})^T$, $J_i \leq J$. Let $\mathbf{y}_i$ be the random response variables, as defined previously. We define another random variable $\bar{\mathbf{y}}$ by $\bar{\mathbf{y}}_i = \text{Diag}(\boldsymbol{\lambda}_i)\mathbf{y}_i$ respectively $\bar{y}_{ij} = \lambda_j y_{ij}$ $(j = 1, \ldots, J_i)$ and apply residual process W_o to $\bar{\mathbf{y}}$. We now apply process W_o to $\bar{\mathbf{y}}$ and process $\mathbf{W}_o$ to $\mathbf{y}$, and to avoid further confusion the superscript will show to which random variable the quantities refer to, for instance $\mathbf{r}^{\bar{\mathbf{y}}}$ refers to $\bar{\mathbf{y}}$ and $\mathbf{r}^{\mathbf{y}}$ to

$\mathbf{y}$.

Note $\mathbf{M}_i^{\bar{\mathbf{y}}} = \mathbf{M}_i^{\mathbf{y}}\text{Diag}(\boldsymbol{\lambda}_i)$, $\mathbf{r}_i^{\bar{\mathbf{y}}} = \text{Diag}(\boldsymbol{\lambda}_i)\mathbf{r}_i^{\mathbf{y}}$ and $\mathbf{A}_i^{\bar{\mathbf{y}}} = \text{Diag}(\boldsymbol{\lambda}_i)\mathbf{A}_i{}^{\mathbf{y}}\text{Diag}(\boldsymbol{\lambda}_i)$. Hence $\mathbf{V}_i^{\bar{\mathbf{y}}} = \text{Diag}(\boldsymbol{\lambda}_i)\mathbf{V}_i{}^{\mathbf{y}}\text{Diag}(\boldsymbol{\lambda}_i)$. It follows that

$$\begin{aligned}
\mathbf{U}_i^{\bar{\mathbf{y}}} &= \mathbf{M}_i^{\bar{\mathbf{y}}}(\mathbf{V}_i{}^{\bar{\mathbf{y}}})^{-1}\mathbf{r}_i^{\bar{\mathbf{y}}} \\
&= \mathbf{M}_i^{\mathbf{y}}\text{Diag}(\boldsymbol{\lambda}_i)\text{Diag}(\boldsymbol{\lambda}_i)^{-1}(\mathbf{V}_i{}^{\mathbf{y}})^{-1}\text{Diag}(\boldsymbol{\lambda}_i)^{-1}\text{Diag}(\boldsymbol{\lambda}_i)\mathbf{r}_i^{\mathbf{y}} \\
&= \mathbf{M}_i^{\mathbf{y}}(\mathbf{V}_i{}^{\mathbf{y}})\mathbf{r}_i^{\mathbf{y}} = \mathbf{U}_i^{\mathbf{y}}
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\Omega}^{\bar{\mathbf{y}}} &= \sum_{i=1}^{n} \mathbf{M}_i^{\bar{\mathbf{y}}}(\mathbf{V}_i{}^{\bar{\mathbf{y}}})^{-1}(\mathbf{M}_i^{\bar{\mathbf{y}}})^T \\
&= \sum_{i=1}^{n} \mathbf{M}_i^{\mathbf{y}}\text{Diag}(\boldsymbol{\lambda}_i)\text{Diag}(\boldsymbol{\lambda}_i)^{-1}(\mathbf{V}_i{}^{\mathbf{y}})^{-1}\text{Diag}(\boldsymbol{\lambda}_i)^{-1}\text{Diag}(\boldsymbol{\lambda}_i)(\mathbf{M}_i^{\bar{\mathbf{y}}})^T \\
&= \sum_{i=1}^{n} \mathbf{M}_i^{\mathbf{y}}(\mathbf{V}_i{}^{\mathbf{y}})^{-1}(\mathbf{M}_i^{\mathbf{y}})^T = \boldsymbol{\Omega}^{\mathbf{y}}.
\end{aligned}$$

This is expected, because different scales should not lead to different GEE estimates $\hat{\boldsymbol{\beta}}$, that is $\mathbf{U}_i$ and $\boldsymbol{\Omega}$ are scale invariant. Also

$$\begin{aligned}
\boldsymbol{\eta}_{\mathbf{W}_o}^{\bar{\mathbf{y}}} &= -n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J_i} \mathbb{1}(\mathbf{t}-\mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t})\mathbf{M}_{ij}^{\bar{\mathbf{y}}} \\
&= -n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J_i} \mathbb{1}(\mathbf{t}-\mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t})\lambda_j\mathbf{M}_{ij}^{\mathbf{y}} \\
&= -n^{-1}\sum_{i=1}^{n} \mathbf{M}_i^{\mathbf{y}}\mathbb{I}(\mathbf{t}-\mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t})\boldsymbol{\lambda}_i \\
&= \left\{-n^{-1}\sum_{i=1}^{n} \mathbb{I}(\mathbf{t}-\mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t})\mathbf{M}_i^{\mathbf{y}}\right\}\boldsymbol{\lambda} = \boldsymbol{\eta}_{W_o}^{\mathbf{y}}\boldsymbol{\lambda}.
\end{aligned}$$

For W_o we obtain:

$$\begin{aligned} W_o^{\bar{\mathbf{y}}} &= n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \mathbb{1}(\mathbf{t} - \mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t}) r_{ij}^{\bar{\mathbf{y}}} \\ &= n^{-1/2} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \mathbb{1}(\mathbf{t} - \mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t}) \lambda_j r_{ij}^{\mathbf{y}} \\ &= n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\lambda}_i^T \mathbb{I}(\mathbf{t} - \mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t}) \mathbf{r}_i^{\mathbf{y}} \\ &= \boldsymbol{\lambda}^T \left\{ n^{-1/2} \sum_{i=1}^{n} \mathbf{I}(\mathbf{t} - \mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t}) \mathbf{r}_i^{\mathbf{y}} \right\} = \boldsymbol{\lambda}^T \mathbf{W}_o^{\mathbf{y}}. \end{aligned}$$

Similarly:

$$\begin{aligned} \widehat{W}_0^{\bar{\mathbf{y}}} &= n^{-1/2} \sum_{i=1}^{n} \left[\sum_{j=1}^{J_i} I(\mathbf{t} - \mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t}) \hat{r}_{ij}^{\bar{\mathbf{y}}} + (\boldsymbol{\eta}_{W_o}^{\bar{\mathbf{y}}})^T (\tilde{\boldsymbol{\Omega}}^{\bar{\mathbf{y}}})^{-1} \tilde{\mathbf{U}}_i^{\bar{\mathbf{y}}} \right] N_i \\ &= n^{-1/2} \sum_{i=1}^{n} \left[\sum_{j=1}^{J_i} I(\mathbf{t} - \mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t}) \lambda_j \hat{r}_{ij}^{\mathbf{y}} + (\boldsymbol{\eta}_{\mathbf{W}_o}^{\mathbf{y}} \boldsymbol{\lambda})^T (\tilde{\boldsymbol{\Omega}}^{\mathbf{y}})^{-1} \tilde{\mathbf{U}}_i^{\mathbf{y}} \right] N_i \\ &= n^{-1/2} \sum_{i=1}^{n} \left[\boldsymbol{\lambda}^T \mathbf{I}(\mathbf{t} - \mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t}) \hat{r}_{ij}^{\mathbf{y}} + \boldsymbol{\lambda}_i^T (\boldsymbol{\eta}_{\mathbf{W}_o}^{\mathbf{y}})^T (\tilde{\boldsymbol{\Omega}}^{\mathbf{y}})^{-1} \tilde{\mathbf{U}}_i^{\mathbf{y}} \right] N_i \\ &= \boldsymbol{\lambda}^T \left\{ n^{-1/2} \sum_{i=1}^{n} \left[\mathbf{I}(\mathbf{t} - \mathbf{b} < \mathbf{z}_{ij} \leq \mathbf{t}) \hat{r}_{ij}^{\mathbf{y}} + (\boldsymbol{\eta}_{\mathbf{W}_o}^{\mathbf{y}})^T (\tilde{\boldsymbol{\Omega}}^{\mathbf{y}})^{-1} \tilde{\mathbf{U}}_i^{\mathbf{y}} \right] N_i \right\} \\ &= \boldsymbol{\lambda}^T \widehat{\mathbf{W}}_0^{\mathbf{y}}. \end{aligned}$$

The processes $W_o^{\bar{\mathbf{y}}}$ and $\widehat{W}_o^{\bar{\mathbf{y}}}$ are asymptotically equivalent by Theorem 7.3.2. We just showed that

$$W_o^{\bar{\mathbf{y}}} \equiv \boldsymbol{\lambda}^T \mathbf{W}_o^{\mathbf{y}} \text{ and } \widehat{W}_o^{\bar{\mathbf{y}}} = \boldsymbol{\lambda}^T \widehat{\mathbf{W}}_o^{\mathbf{y}}, \tag{7.14}$$

hence $\boldsymbol{\lambda}^T \mathbf{W}_o$ and $\boldsymbol{\lambda}^T \widehat{\mathbf{W}}_o$ are also asymptotically equivalent. It follows now from Theorem 7.3.8 that $\mathbf{W}_o$ and $\widehat{\mathbf{W}}_o$ are also asymptotically equivalent. $\mathbf{W}_k$ is a subcase of $\mathbf{W}_o$ and it follows that $\mathbf{W}_k$ and $\widehat{\mathbf{W}}_k$ are asymptotically equivalent. In a similar manner, this can also be shown for $\mathbf{W}_p$ and $\widehat{\mathbf{W}}_p$. The asymptotic equiva-

lence of $h(\mathbf{W})$ and $h(\hat{\mathbf{W}})$ follows from the continuous mapping theorem.

Consistency of the supremum tests:

The consistency of similar supremum tests was shown/mentioned in several papers (Su and Wei 1991, Lin et al. 1993, Lin et al. 2002, Pan and Lin 2005, Arbogast and Lin 2005). It was shown or mentioned that under certain sufficient conditions $n^{-1/2}W_o(\mathbf{t}_0;\hat{\boldsymbol{\beta}}) \rightarrow_p J \neq 0$ for at least some $\mathbf{t}_0$, hence, $n^{-1/2}G_{W_0}$ converges to a nonzero constant.

We want to show now the consistency of $G_{h(\mathbf{W}_o)}$. First, we show that $n^{-1/2}(\mathbf{W}_o)_j$ converges to a non-zero constant J_j. As before we use (7.14) and set $\boldsymbol{\lambda} := \mathbf{e}_j$, where $\mathbf{e}_j$ is the jth unit vector. We have now $W_o \equiv \mathbf{e}_j^T\mathbf{W}_o = (\mathbf{W}_o)_j$. From the above, we can conclude $n^{-1/2}W_o \rightarrow_p J_j \neq 0$, or equivalently $n^{-1/2}(\mathbf{W}_o)_j \rightarrow_p J_j \neq 0$, that is the consistency of $G_{\mathbf{W}_o}$.

To show that the test $G_{h(\mathbf{W}_o)}$ is consistent, it is sufficient to show $n^{-1/2}\ h(\mathbf{W}_o)$ converges to a nonzero vector for some $\mathbf{t}_0$ (then $n^{-1/2}G_{h(\mathbf{W}_o)}$ converges to a nonzero constant). We just established $n^{-1/2}\mathbf{W}_o \rightarrow_p \mathbf{c}$ with $\mathbf{c}$ being nonzero in all components. Thus, $n^{-1/2}h(\mathbf{W}_o) \rightarrow_p h(\mathbf{c})$. We have $\mathbf{0} < |\mathbf{c}|$ and it follows from the monotonicity condition $\mathbf{0} = |h(\mathbf{0})| < |h(\mathbf{c})|$, which was to be shown.

Similarly we proceed with $\mathbf{W}_p$ and $\mathbf{W}_g$. □

Remark 7.3.9. $\mathcal{J} = \mathbb{E}\mathcal{I}$ and $\mathcal{I}$ are asymptotically equivalent. For a MGLM, we can also use $\mathcal{I}$ instead of $\mathcal{J}$ in the definition of $\boldsymbol{\Omega}$ in (7.8). The resulting cumulative residual processes are still asymptotically equivalent. Arbogast and Lin (2005), who considered the cumulative residual processes for logistic regression applying ML methodology, used the observed information matrix $\mathcal{I}$ in the definition of $\widehat{W}$.

7.3.2 Residual Processes for the Proportional Odds Model

We now express the proportional odds model (7.1) as a multivariate generalised linear model (MGLM), see Subsection 5.2.2 on page 155 for details. The probability $\pi_{ij} = P(Y = j \mid \mathbf{x}_i) \equiv \mu_{ij}$ can be computed from the cumulative probabilities $\pi^*_{ij} = \pi_{i1} + \pi_{i2} + \cdots + \pi_{ij} = P(Y \leq j \mid \mathbf{x}_i)$ by $\pi_{ij} = \pi^*_{ij} - \pi^*_{i,j-1}$ for $j > 1$ and $\pi_{i1} = \pi^*_{i1}$. We have $\boldsymbol{\Sigma}_i = \text{Diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i^T$ and the proportional odds model can be re-expressed as

$$g_j(\boldsymbol{\mu}_i) = g_j(\boldsymbol{\pi}_i) = \log\left(\frac{\pi^*_{ij}}{1 - \pi^*_{ij}}\right) = \alpha_j + \mathbf{x}_i^T \boldsymbol{\gamma}, j = 1, ..., J-1$$

or in more complex form of a MGLM

$$\mathbf{g}(\boldsymbol{\mu}_i) = \mathbf{Z}_i \boldsymbol{\beta}$$

with $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\gamma}^T)^T$, $\mathbf{g} = (g_1, \ldots, g_{J-1})^T$, and

$$\mathbf{Z}_i = \begin{pmatrix} 1 & & & \mathbf{x}_i^T \\ & 1 & & \mathbf{x}_i^T \\ & & \ddots & \vdots \\ & & 1 & \mathbf{x}_i^T \end{pmatrix} = \left(\mathbf{I}_q, \mathbf{1}_{q \times 1} \otimes \mathbf{x}_i^T\right), \tag{7.15}$$

where $\mathbf{I}_q \in \mathbb{R}^{q \times q}$ is the identity matrix and $\mathbf{1}_{a \times b} \in \mathbb{R}^{a \times b}$ is the matrix containing only 1's. See also Fahrmeir and Tutz (2001, pp. 81-98) for cumulative models and expressing them as MGLM. For completeness, we give compact formulae for the first and second derivatives of the log-likelihood for the proportional odds model,

which are needed to apply the previously introduced stochastic processes. Note

$$\frac{\partial \boldsymbol{\pi}_i^T}{\partial \boldsymbol{\pi}_i^*} = \begin{pmatrix} 1 & -1 & & \\ & 1 & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{pmatrix} \in \mathbb{R}^{q \times q} \qquad \frac{\partial (\boldsymbol{\pi}_i^*)^T}{\partial \boldsymbol{\pi}_i} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ & 1 & \dots & 1 \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix} \in \mathbb{R}^{q \times q}.$$

Also

$$\mathbf{M}_i = \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} = \mathbf{Z}_i^T \text{Diag}\left\{\boldsymbol{\pi}_i^*(\mathbf{1}_{q\times 1} - \boldsymbol{\pi}_i^*)\right\} \frac{\partial \boldsymbol{\pi}_i^T}{\partial \boldsymbol{\pi}_i^*}$$

and

$$\begin{aligned} \mathcal{I} =& -\frac{\partial l^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \\ =& \mathbf{U}\mathbf{U}^T - \sum_{i=1}^{n} \mathbf{Z}_i^T \text{Diag}[\text{Diag}\{(\mathbf{1}_{q\times 1} - 2\boldsymbol{\pi}_i^*)\boldsymbol{\pi}_i^*(\mathbf{1}_{q\times 1} - \boldsymbol{\pi}_i^*)\} \frac{\partial \boldsymbol{\pi}_i^T}{\partial \boldsymbol{\pi}_i^*} \boldsymbol{\Sigma}^{-1} \mathbf{r}_i] \mathbf{Z}_i \end{aligned}$$

with $\mathbf{r}_i = \mathbf{y}_i - \boldsymbol{\pi}_i$.

Let us now focus on the process $\mathbf{W}_k$, which only checks the functional form of the kth covariate. We use $\mathbf{U}_i = \partial l_i / \partial \boldsymbol{\beta} \equiv \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$ and $\boldsymbol{\Omega} = n^{-1}\mathcal{I}$ for the computation of the processes similarly defined as (7.10) and (7.11) and let these processes be denoted by $\mathbf{W}_k^m$ and $\widehat{\mathbf{W}}_k^m$, where m stands for the multinomial residuals $\mathbf{r}_i$.

Instead of using the multivariate residuals $\mathbf{r}_i$, we can also use the multivariate cumulative residuals $\mathbf{r}_i^*$ defined by

$$\mathbf{r}_i^* = \mathbf{y}_i^* - \boldsymbol{\pi}_i^*,$$

where $\mathbf{r}_i^* = (r_{i1}^*, r_{i2}^*, \dots, r_{i(J-1)}^*)^T$, $\mathbf{y}_i^* = (y_{i1}^*, y_{i2}^*, \dots, y_{i(J-1)}^*)^T$, and $\boldsymbol{\pi}_i^* = (P(Y \leq 1|\mathbf{x}_i), P(Y \leq 2|\mathbf{x}_i), \dots, P(Y \leq J-1|\mathbf{x_i}))^T$. Section 7.2 defined the notations r_{ij}^*

and y_{ij}^*. We consider the multivariate stochastic process

$$\mathbf{W}_k^*(t;\hat{\boldsymbol{\beta}}) = n^{-1/2}\sum_{i=1}^{n}\mathbb{1}(x_{ik}\leq t)\hat{\mathbf{r}}_i^*.$$

Similarly, if the model holds, $\mathbf{W}_k^*(t;\hat{\boldsymbol{\beta}})$ converges weakly to a vector of zero-mean Gaussian processes, because $\mathbf{y}^*$ can also be considered as observations and the GEE methodology and Theorem 7.3.3 applies. The distribution of the processes can be approximated by $\widehat{\mathbf{W}}_k^*(t;\hat{\boldsymbol{\beta}})$, which has the same form as $\widehat{\mathbf{W}}_k^m(t;\hat{\boldsymbol{\beta}})$ but replacing $\hat{\mathbf{r}}_i$ with $\hat{\mathbf{r}}_i^*$ and in $\boldsymbol{\eta}$ replacing $\hat{\boldsymbol{\pi}}_i$ with $\hat{\boldsymbol{\pi}}_i^*$. Table 7.2 gives a summary of all graphical diagnostic methods for the two approaches for the proportional odds model.

Table 7.2: Notations used for graphical diagnostic methods

Notation	Approach	Description
B_j	Binary	Collapse the response categories into $(\leq j, > j)$
Bonf(B)	Binary	Bonferroni: compare the p-value with $\alpha/(J-1)$
$(\mathbf{W}^m)_j$	Mult ($\mathbf{r}$)	Using the jth component of residual $\mathbf{r}$
Bonf($\mathbf{W}^m$)	Mult ($\mathbf{r}$)	Bonferroni: compare the p-value with $\alpha/(J-1)$
sum($\mathbf{W}^m$)	Mult ($\mathbf{r}$)	Using function sum($\mathbf{W}^m$) := $\sum_{j=1}^{J-1}(W^m)_j$
prod($\mathbf{W}^m$)	Mult ($\mathbf{r}$)	Using function prod($\mathbf{W}^m$) := $\prod_{j=1}^{J-1}(W^m)_j$
max($\mathbf{W}^m$)	Mult ($\mathbf{r}$)	Using function max($\mathbf{W}^m$) := max$\lvert\mathbf{W}^m\rvert$
$(\mathbf{W}^*)_j$	Mult ($\mathbf{r}^*$)	Using the jth component of residual $\mathbf{r}^*$
Bonf($\mathbf{W}^*$)	Mult ($\mathbf{r}^*$)	Bonferroni: compare the p-value with $\alpha/(J-1)$
sum($\mathbf{W}^*$)	Mult ($\mathbf{r}^*$)	Using function sum($\mathbf{W}^*$) := $\sum_{j=1}^{J-1}(W^*)_j$
prod($\mathbf{W}^*$)	Mult ($\mathbf{r}^*$)	Using function prod($\mathbf{W}^*$) := $\prod_{j=1}^{J-1}(W^*)_j$
max($\mathbf{W}^*$)	Mult ($\mathbf{r}^*$)	Using function max($\mathbf{W}^*$) := max$\lvert\mathbf{W}^*\rvert$

Mult ... Multivariate

Remark 7.3.10. Let us extend the dimension of the observations by 1 to $\mathbf{y}_i = (y_{i1},\ldots,y_{iJ})^T$ with $y_{iJ} = 1$ if response $Y_i = J$. In the proof of Theorem 7.3.3, we consider the univariate process $\boldsymbol{\lambda}^T\mathbf{W}$. This univariate process can be thought of as a process of observations $\boldsymbol{\lambda}^T\mathbf{y}_i$. First, we note from the same proof it follows

that the process $\mathbf{W}$ is linear in $\mathbf{y}_i$. Now we regard $\boldsymbol{\lambda}$ as a score vector, assigning a score λ_j for each response j, for example equally spaced integer scores $\lambda_j = j$ considered by Lipsitz et al. (1996). The process $\boldsymbol{\lambda}^T\mathbf{W}$ is then identical to $-\text{sum}(\mathbf{W}^*)$. Lipsitz et al. (1996) ($j = 1$) considered another option $\lambda_j = 1$ and $s_{j'} = 0$ for $j' \neq j$. Then $\boldsymbol{\lambda}^T\mathbf{W}$ is identical to the jth component of $\mathbf{W}$. They used the scores to project the multivariate mean vector to a univariate mean, called the *mean score* computed by $\boldsymbol{\lambda}^T\boldsymbol{\mu}_i$ for some reasonable choice of scores $\boldsymbol{\lambda}$.

Because of the two mentioned equivalences, it is sufficient to consider $\mathbf{W}$, $\mathbf{W}^*$ and the above mentioned functions/projections. Also note, that $(\mathbf{W}^*)_j = \boldsymbol{\lambda}_j^T\mathbf{W}^m$ for $\boldsymbol{\lambda}_j^T = (\mathbf{1}_j, \mathbf{0}_{J-1-j})$. Hence, we can compute $\mathbf{W}^*$ by the linear transformation $\mathbf{W}^* = \boldsymbol{\Lambda}\mathbf{W}^m$ with matrix $\boldsymbol{\Lambda}$ containing such rows $\boldsymbol{\lambda}_j^T$, similarly $\widehat{\mathbf{W}}^*$.

7.3.3 Comments about the Computation of the Gaussian Processes

Given the parameter estimates for the data, the computation of the $\mathbf{W}_k$'s is relatively easy. The vector of residuals $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)^T$ is a by-product of the fitting and the computation of the $\mathbf{W}_k$'s only requires the computation of the so far unknown indicator functions $\mathbb{1}(x_{ik} \leq t)$. We do not need to compute $\mathbb{1}(x_{ik} \leq t)$ for infinite many t, but only for the number $m \leq n$ of different values $t_1, \ldots, t_m$ for the kth covariate. We can store all these $\mathbb{1}(x_{ik} \leq t)$ in an $n \times m$ matrix $\mathbb{I}(\mathbf{x}_k)$. For given $\mathbf{r}$ and $\mathbb{I}(\mathbf{x}_k)$, the computation of $\mathbf{W}_k$ requires simple matrix operations.

The computation of the $\widehat{\mathbf{W}}_k$'s is much more laborious, because we need to resample a large number ($M \geq 1000$) of realisations of the $\widehat{\mathbf{W}}_k$'s. As a by-product from the fitting algorithm we obtain $\boldsymbol{\Omega}$, $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_n)^T$ and $\mathbf{M} = (\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_n)^T$.

From $\mathbb{I}(\mathbf{x}_k)$ and $\mathbf{M}$ we can compute $\boldsymbol{\eta}(t_1), \ldots, \boldsymbol{\eta}(t_m)$. In the definition of the

$\widehat{\mathbf{W}}_k$'s, which have the form $\sum_{i=1}^{n}[\dots]_i N_i$, the quantities in the bracket terms $[\dots]_i$ can be computed by matrix operations and can be stored in an $n \times (J-1) \times m$ array $\mathbf{B}$. Now we generate M times the n realisations $N_1, \dots, N_n$ from $N(0,1)$ and then store them in the $M \times n$ matrix $\mathbf{N}$. Finally, we can compute the $\widehat{\mathbf{W}}_k$'s from $\mathbf{N}$ and $\mathbf{B}$ by M matrix multiplications. Or we apply a tensor product to reduce computation time further by avoiding the M matrix multiplications, because in many computer languages, such as R or Matlab, the summation of products using loops takes much longer than using matrix/tensor multiplication instead. Also note that $\mathbf{W}_k^* = \mathbf{\Lambda}\mathbf{W}_k^m$ and similarly $\widehat{\mathbf{W}}_k^* = \mathbf{\Lambda}\widehat{\mathbf{W}}_k^m$, see Remark 7.3.10. In fact, for the multivariate approach, we only need to compute $\mathbf{W}^m$ and the $\widehat{\mathbf{W}}^m$'s. Given these processes, all other processes can be relatively easily computed. We conclude, an efficient implementation of the cumulative residual processes and their approximation is essential in yielding a fast computational routine.

7.4 Simulation Study

We proposed two approaches including 9 graphical diagnostic methods to detect model inadequacy in the proportional odds model. To compare the performances of these methods, we undertake a small-scale simulation study to investigate the power under a fixed alternative $\mathcal{H}_1$ and the Type I error rate under $\mathcal{H}_0$. We investigate two forms of functional mis-specification in a single covariate x. We consider discrete x in one scenario and continuous in the other. For each situation, the empirical Type I error rate and powers are estimated based on the proportion of rejected null hypotheses in 10,000 simulated datasets.

Scenario 1:

Let $J = 3$. We consider the true model as follows:

$$\text{logit}[P(Y \leq j \mid X)] = \alpha_j - \gamma_1 X - \gamma_2 X^2,\ j = 1,\ 2. \tag{7.16}$$

We first generate grouped categorical X observations with values ranging from -5 to +5 with equal probability, representing a discrete uniform distribution. Conditional on the X-values Y values are generated from model (7.16) by choosing $\alpha_1 = -2$, $\alpha_2 = -1$, $\gamma_1 = +0.25$, and $\gamma_2 = 0.0, -0.05, -0.1$, and then simulating multinomial random variables with three categories. We generate 110 observations in each dataset, rendering approximately 10 occurrences for each distinct X-value on an average.

We try to fit a simple model with just the linear term to the simulated data with X^2 omitted, namely,

$$\text{logit}[P(Y \leq j \mid X)] = \alpha_j - \gamma_1 X,\ j = 1,\ 2. \tag{7.17}$$

When $\gamma_2 = 0.0$, the model is correctly specified and we can estimate the rejection rate under this $\mathcal{H}_0$ and compare this estimate of Type I error rate with the significance level (α), which was always set at 0.05. When $\gamma_2 = -0.05$, or -0.1, we evaluate the performance of the different graphical diagnostic methods by their power to detect departures from the correct model. Table 7.3 summarises the results for this scenario in the first 3 columns. Among all the methods compared, the naive binary collapsing approach exhibits the worst performance. It fails to maintain the nominal Type I error level and the estimated Type I error rate is twice the desired level of significance $\alpha\,(= 0.05)$. The multivariate approaches

based on the residuals and the cumulative residuals produce better results. Both of the multivariate residuals ($\mathbf{r}$) and multivariate cumulative residuals ($\mathbf{r}^*$) maintain the correct level of significance under a correctly specified model with $\gamma_2 = 0$. The power for the multivariate methods based on the functionals sum($\mathbf{W}^m$) and sum ($\mathbf{W}^*$) appears to be the best.

Table 7.3: Simulation results for scenarios 1 and 2 showing the power under $\mathcal{H}_0$ and $\mathcal{H}_1$ for various values of γ

	The functional form of $\mathbf{x}\boldsymbol{\gamma}$ for the true model					
	$0.25X + \gamma X^2$			$\gamma\cos(X)$		
Methods	$\gamma = 0.00$	$\gamma = -0.05$	$\gamma = -0.10$	$\gamma = 0.0$	$\gamma = -1.0$	$\gamma = -3.0$
B_1	0.148	0.435	0.934	0.168	0.393	0.981
B_2	0.097	0.482	0.959	0.155	0.407	0.822
Bonf(B)	0.126	0.491	0.964	0.180	0.433	0.969
Bonf($\mathbf{W}^m$)	0.042	0.220	0.811	0.046	0.129	0.879
sum($\mathbf{W}^m$)	0.051	0.285	0.855	0.052	0.179	0.591
prod($\mathbf{W}^m$)	0.054	0.113	0.386	0.058	0.112	0.704
max($\mathbf{W}^m$)	0.035	0.102	0.543	0.056	0.086	0.836
Bonf($\mathbf{W}^*$)	0.043	0.292	0.895	0.049	0.191	0.906
sum($\mathbf{W}^*$)	0.048	0.357	0.947	0.049	0.344	0.974
prod($\mathbf{W}^*$)	0.047	0.340	0.941	0.049	0.270	0.958
max($\mathbf{W}^*$)	0.041	0.266	0.874	0.051	0.203	0.939
Wald-$\gamma = 0$	0.050	0.568	0.994	-	-	-
HL (G=5)	0.049	0.278	0.894	0.046	0.255	0.949

Scenario 2

The second scenario represents a situation where the cumulative logit probabilities associated with the response are related in a non-linear manner with X, but are linear in $\cos(X)$. The correct model is as follows

$$\text{logit}(\Pr(Y \leq j \mid X)) = \alpha_j - \gamma \cos X,\ j = 1,\ 2, \tag{7.18}$$

where $\alpha_1 = -1$, $\alpha_2 = 1$, and $\gamma = 0, -1, -3$. We simulated X from a standard normal distribution and conditional on X simulated Y from the multinomial distribution with probabilities defined using (7.18). Again we fit each simulated dataset using the model (7.17) with a linear term of X. Table 7.3 summarises the results in the last 3 columns. Similar to the first scenario, the binary collapsing approach gives a overly liberal result that rejects the null hypothesis more often than we expect and consequently has inflated power values. Among the methods in the multivariate approach, the sum($\mathbf{W}^*$) has the best performance in terms of maintaining Type I error and attaining higher power values.

A goodness-of-fit statistic as proposed in Lipsitz et al. (1996) based on the mean score is also included in the simulation study for comparison purposes. According to the percentiles of the predicted mean score, subjects are partitioned into G regions as defined in Lipsitz et al. (1996). Given the partition of the data, the following model is fitted

$$\text{logit}[\Pr(Y \leq j \mid \mathbf{x})] = \alpha_j - \mathbf{x}\boldsymbol{\gamma} + \sum_{g=1}^{G-1} \mathbb{1}_{ig}\delta_g \tag{7.19}$$

where $\mathbb{1}_{ig}$ are group indicators with $\mathbb{1}_{ig} = 1$ if $\boldsymbol{\lambda}^T\hat{\boldsymbol{\pi}}_i$ is in region g and $\mathbb{1}_{ig} = 0$ otherwise, for equally spaced integer scores $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_J)^T$ with $\lambda_j = j$, see also Remark 7.3.10 on page 254. If model (7.17) is correct, then $\delta_1 = \delta_2 = \cdots = \delta_{G-1} = 0$ independently of the chosen regions and scores. We simply test $\mathcal{H}_0$: $\delta_1 = \delta_2 = \cdots = \delta_{G-1} = 0$ and compute a likelihood-ratio (LR), Wald and a score statistic. We refer to this statistic as Hosmer-Lemeshow (HL)-type statistic, because the idea stems from the HL statistic developed for logistic regression as extended to ordinal responses. The LR test, the Wald-test, and the score test in this case are asymptotically equivalent and showed quite similar power values;

hence, Table 7.3 lists only the result of the HL-type score tests.

For the first scenario, Table 7.3 also gives the Wald test on the null hypothesis $\mathcal{H}_0$: $\gamma_2 = 0$. If we do know that the correct model includes the X^2 term, this test is optimal as one would expect, but the Wald test is not applicable when the true functional form is unknown. Thus in situation 2, we cannot formulate an appropriate Wald test to compare the two models in terms of a single parameter.

Summary of Simulation Results

In general, the graphical diagnostic methods sum($\mathbf{W}$) and prod($\mathbf{W}$) have good power properties. We do expect the graphical diagnostic methods to provide a lower power compared with the Wald test when the true model contains the term X^2 as in Scenario 1. Unlike the Wald test, the graphical diagnostic methods do not focus on any specific term. It checks model mis-specification for a wide range of the mis-specification in a non-parametric manner (e.g. the functional form could be anything like X^2, $\log X$, X^3, $\cos X$, etc). Arbogast and Lin (2005) also pointed out that the Wald test cannot be used to check whether the chosen functional term is satisfactory. Remarkably, some of the graphical diagnostic methods are very comparable with the optimal Wald test in terms of power for Scenario 1, when one is testing for the missing term in the true model, with a true model known. For example, the sum($\mathbf{W}^*$) gives a power of 0.947 when the true coefficient of X^2 is 0.10. The Wald test gives a power of 0.994 in comparison. On the other hand, the graphical methods of "Bonf", "sum" and "prod" using the cumulative residuals ($\mathbf{r}^*$) in the multivariate approach have higher power than the overall Hosmer-Lemeshow test in scenario 1. The methods with 'sum', and 'prod' using the cumulative residuals ($\mathbf{r}^*$) still give higher power than the overall HL test in Scenario 2. The diagnostic based on sum($\mathbf{W}^*$) appears to the best choice based in

our limited simulation settings.

7.5 Examples

In the following, we illustrate the methods of Sections 7.2 and 7.3 using the two examples mentioned previously.

7.5.1 Yield of New Hybrid Tomato

To illustrate the methods, we first fit the proportional odds model (7.1) to the data given in Table 7.1. An agronomist studied the effects of moisture (X_1, in inches) and temperature(X_2, in ${}^0C$) on the yield of a new hybrid tomato (Y). The model includes all main effects of the covariates X_1 and X_2. The coefficient for *Moist* is 0.7418 (with s.e. of 0.2355) and the coefficient for *Temp* is 0.5348 (with s.e. of 0.3299), see Table 7.4. The moisture effects are significant and temperature effects

Table 7.4: Yield of new Hybrid Tomato

Model	Predictor	Coef	S.E.	Wald Z	P-value
Model 1	Temp	0.5348	0.3299	-1.62	0.1050
	Moist	0.7418	0.2355	-3.15	0.0016
Model 2	Temp	1.4478	0.7462	-1.94	0.0524
	Moist	-11.9012	5.6269	2.12	0.0344
	Moist^2	0.7212	0.3415	-2.11	0.0347

are moderately significant. The description of the fitted model is that, given the temperature level is fixed, the odds of having higher yield of a new hybrid tomato are estimated to be $e^{0.7418} = 2.1$ times higher for a one inch decrease in moisture level.

Table 7.5: The p-values of testing model mis-specification based on graphical diagnostics

Tests	*Temperature*	*Moisture*	α (Bonferroni adjustment)
B_1	0.0845	0.0016	0.05 (0.025)
B_2	0.0383	0.0001	0.05 (0.025)
$(\mathbf{W}^m)_1$	0.2216	0.0070	0.05 (0.025)
$(\mathbf{W}^m)_2$	0.1984	0.0094	0.05 (0.025)
$(\mathbf{W}^*)_1$	0.2216	0.0070	0.05 (0.025)
$(\mathbf{W}^*)_2$	0.2654	0.0011	0.05 (0.025)
sum$(\mathbf{W}^m)$	0.2654	0.0011	0.05
max$(\mathbf{W}^m)$	0.2520	0.0157	0.05
prod$(\mathbf{W}^m)$	0.1874	0.0084	0.05
sum$(\mathbf{W}^*)$	0.6927	0.0060	0.05
max$(\mathbf{W}^*)$	0.2682	0.0074	0.05
prod$(\mathbf{W}^*)$	0.2038	0.0423	0.05

We used different plots to check the model mis-specification for *Temp* and *Moist*. Table 7.5 shows the p-values for each plot/process. Figure 7.1 gives the plot using the method $(\mathbf{W}^m)_2$ for *Moist*. The dark black dashed line indicates the observed process and the fine solid lines indicate the simulated realisations. The p-value of testing that the model has a correct functional form in *Moist* is 0.0094. Figure 7.2 gives the plot using the method sum$(\mathbf{W}^*)$, with p-value of 0.0060. The results suggest that there is model mis-specification for the proportional odds model with the covariate *Moist*, but not with the covariate *Temp* disregarding the unreliable method B_2.

We re-fit the proportional odds model including the higher-order term for X_1 (*Moist*). All of the coefficients for *Temp*, *Moist*, and *Moist*^{2} are significant with p-values 0.05, 0.03, and 0.03, respectively, see Table 7.4. Also, the coefficients are 1.4578, -11.9012, and 0.7213, respectively. Therefore, the logit of the cumulative probability in Y does not have a linear relationship with the moisture level. The yield of a new hybrid tomato increases when the moisture level increases to 8

Figure 7.1: Plot of residuals against *Moist* using the method $(\mathbf{W}^m)_2$ to check the model mis-specification for the model of *Temp* + *Moist*. The dark black line indicates the observed process and the fine lines indicate the simulated realisations.

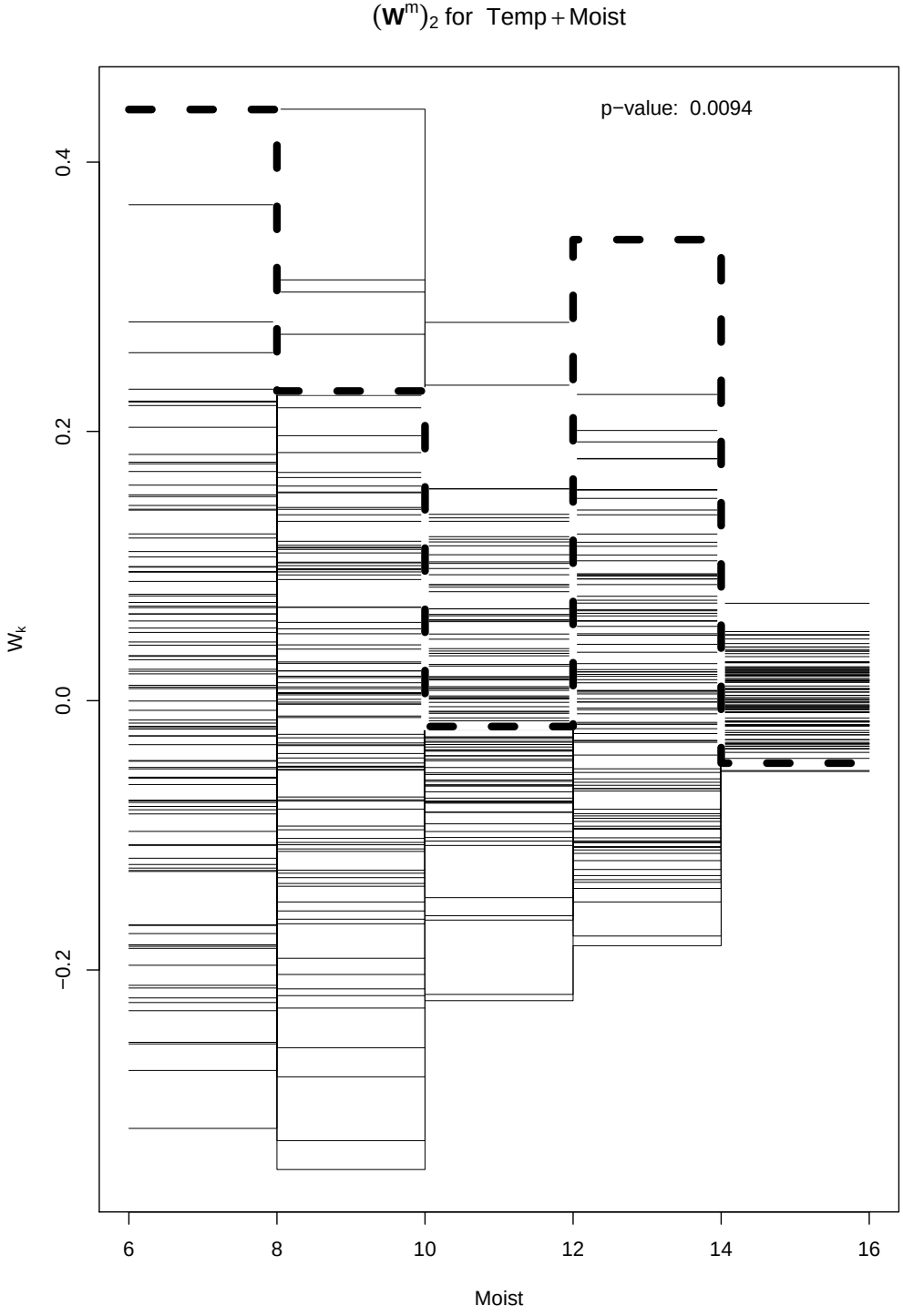
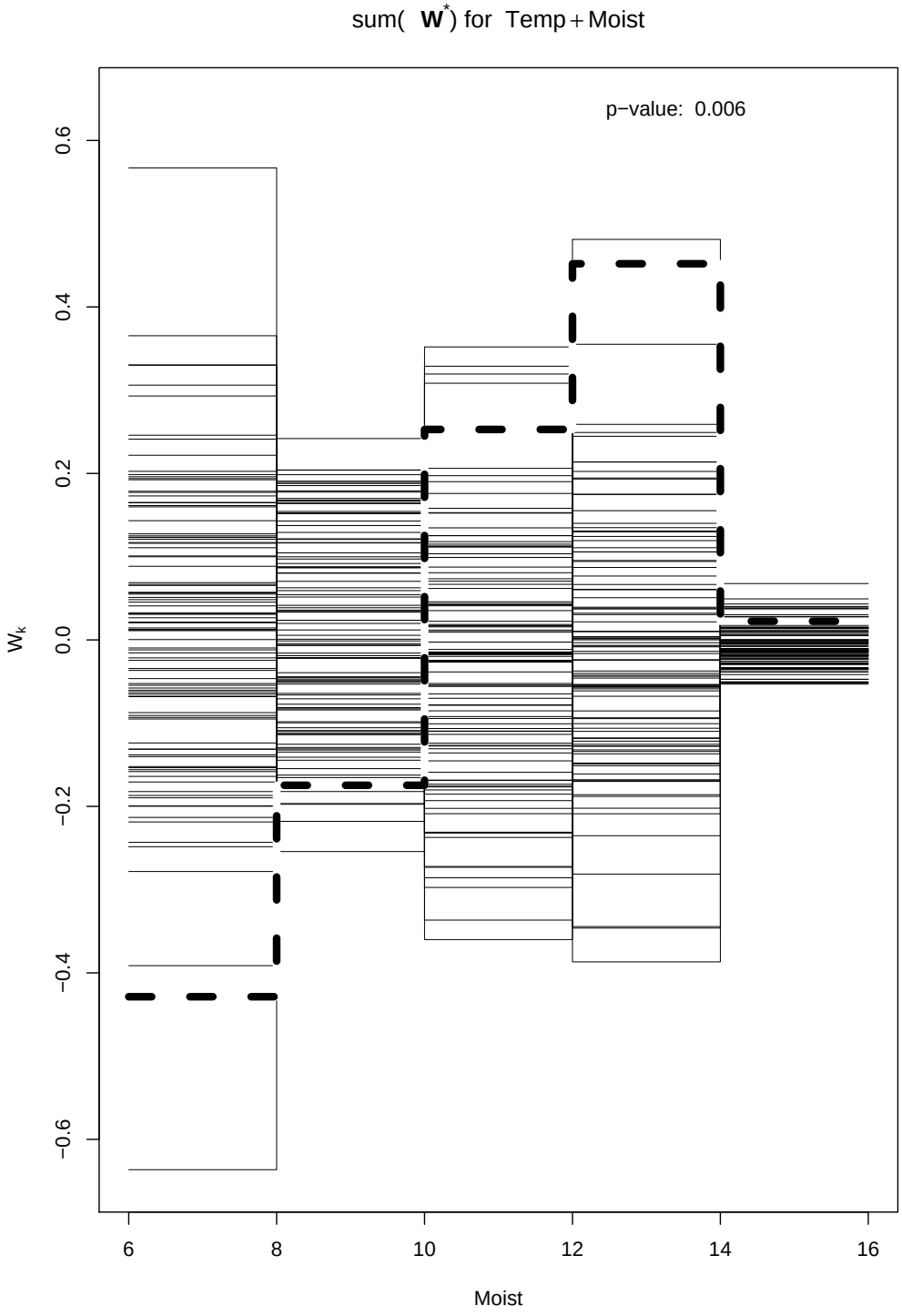


Figure 7.2: Plot of residuals against *Moist* using the method sum($\mathbf{W}^*$) to check the model mis-specification for the model of *Temp* + *Moist*. The dark black line indicates the observed process and the fine lines indicate the simulated realisations.



inches, and then the yield decreases when the moisture level increases from 8 to 14 inches.

Table 7.6: The p-values of testing model mis-specification based on graphical diagnostics for model Temp+Moist+Moist^2

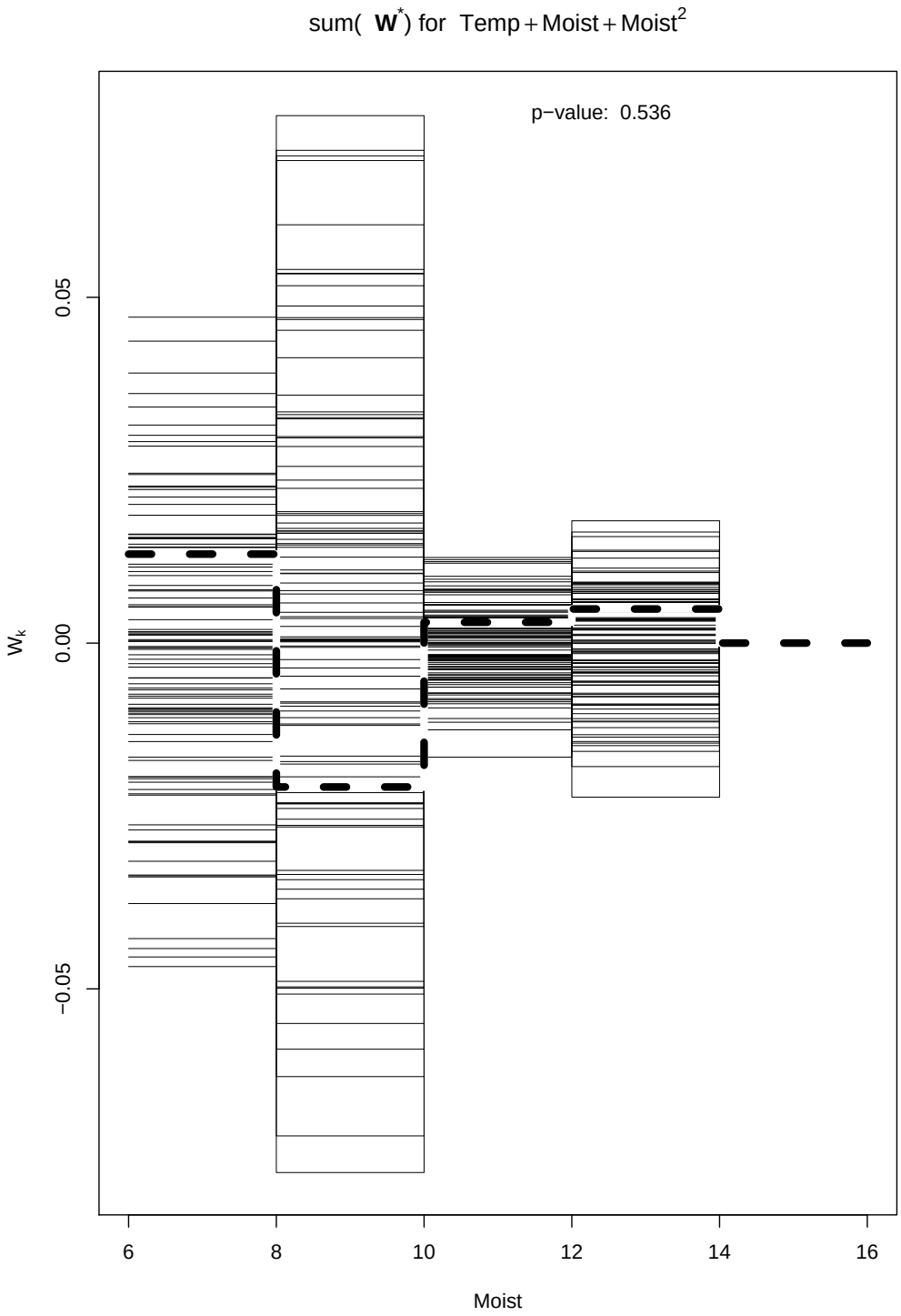
Tests	*Temperature*	*Moisture*	α (Bonferroni adjustment)
B_1	0.2387	0.2680	0.05 (0.025)
B_2	0.0001	0.0001	0.05 (0.025)
$(\mathbf{W}^m)_1$	0.2567	0.5421	0.05 (0.025)
$(\mathbf{W}^m)_2$	0.2560	0.5423	0.05 (0.025)
$(\mathbf{W}^*)_1$	0.2567	0.5421	0.05 (0.025)
$(\mathbf{W}^*)_2$	0.5818	0.5539	0.05 (0.025)
sum$(\mathbf{W}^m)$	0.5818	0.5539	0.05
max$(\mathbf{W}^m)$	0.2609	0.5423	0.05
prod$(\mathbf{W}^m)$	0.2550	0.5422	0.05
sum$(\mathbf{W}^*)$	0.2666	0.5360	0.05
max$(\mathbf{W}^*)$	0.2567	0.5421	0.05
prod$(\mathbf{W}^*)$	0.4258	0.6460	0.05

Figure 7.3 shows the plot using the method sum$(\mathbf{W}^*)$ for the new model. It gives the p-value of 0.536. Table 7.6 shows the p-values for all introduced methods. All methods except the unreliable B_2 do not show model mis-specification for the new model. The functional terms chosen in the final model are satisfactory.

7.5.2 Normative Aging Study

The Normative Aging Study (NAS) is a multidisciplinary longitudinal study of aging in men established by the Veteran's Administration of the United States in 1963. NAS subjects have reported for medical examination every 3 to 5 years. Though the study records data on a wide spectrum of variables, including several health related measures, dietary and behavioural exposures, exposure to certain metals in their environment, and psychosocial events, our analysis focuses on exploring the relationship of fasting blood glucose (FBG), the level of glucose,

Figure 7.3: Plot of residuals against *Moist* using the method sum($\mathbf{W}^*$) to check the model mis-specification for the model of *Temp* + *Moist* + *Moist*^{2}. The dark black line indicates the observed process and the fine lines indicate the simulated realisations.



with two markers of systemic inflammation, namely, white blood cell count (*wbc*) and blood levels of C-reactive protein (*crp*) after controlling for age and smoking status. The measurements were taken during January 2000 to December 2004 and we consider only the last complete observation available on the subject in case multiple measurements were available on the same subject.

The current dataset as shown in Table E on page 313 contains observations on 682 men in the age range of 48 to 93 years. FBG was categorised into three categories according to the clinical definition of diabetes (The Expert Committee 1997), with FBG $<$ 110mg/dl termed as normal (category 1), between 110 and 126 mg/dl termed as impaired fasting glucose (category 2) and $\geq$ 126mg/dl termed as diabetes (category 3). It has been suggested in the literature that oxidative stress-induced inflammatory response increases insulin resistance, resulting in hyperglycemia or elevated levels of FBG which in turn causes oxidative stress again (Pliquett et al. 2004). Inflammation is known to be a risk factor for diabetes (Nakanish et al. 2003). White blood cell count and C-reactive protein can be viewed as biomarkers of systemic inflammation and thus could potentially be associated with FBG levels, leading to this analysis.

We first try to fit a simple model that includes linear terms of the covariates *wbc*, *crp*, *age*, and *smoking*. In this analysis the effect of *wbc* on FBG turns out to be marginally significant with p-value 0.0857 with fitted estimate of β as 0.041; *crp* is not significant with p-value 0.27 and fitted estimate of β as 0.094 (see Table 7.7). The interpretation of the fitted model, for example, in terms of the *wbc* effect is that given fixed values of all other covariates in the model, the odds of having fasting blood glucose towards higher end of the FBG scale with one unit increase in WBC are estimated to be $e^{0.041}$ or 1.04 times higher than having values on the lower end of the FBG scale. Neither age nor smoking status was found to be asso-

Table 7.7: Parameter estimates and p-values for the fitted proportional odds model using covariates "age+smk+wbc+crp" (Model 1) followed by the model "age+smk+wbc+wbc^2+wbc^3+crp+crp^2" (Model 2) in the Normative Aging Study Example.

Model	Predictor	Coef	S.E.	Wald Z	P-value
Model 1	age	-0.00747	0.01255	-0.60	0.5516
	smk	0.03331	0.06186	0.54	0.5902
	wbc	0.04134	0.02406	1.72	0.0857
	crp	0.09408	0.08572	1.10	0.2724
Model 2	age	-0.0080846	0.0126358	-0.64	0.5223
	smk	0.0442334	0.0624535	0.71	0.4788
	wbc	0.5628662	0.2464199	2.28	0.0224
	wbc^2	-0.0376317	0.0192671	-1.95	0.0508
	wbc^3	0.0005244	0.0002956	1.77	0.0760
	crp	0.3960383	0.1821148	2.17	0.0297
	crp^2	-0.0297128	0.0198322	-1.50	0.1341

ciated with FBG levels. Hence there appears to be a positive association between FBG and *wbc* and *crp*, but none of them are statistically significant.

We used different diagnostic tools to check the model mis-specification for *age*, *smoking*, *wbc* and *crp*. Table 7.8 presents the p-value corresponding to each of the graphical methods. Figure 7.4 gives the plot using the method $(\mathbf{W}^m)_1$ for *wbc*, whereas Figure 7.5 shows the same for *crp*. The dark black dashed line indicates the observed process and the fine solid lines indicate the simulated realisations. We calculate the p-value using 1000 simulated realisations, while the figure only shows 100 of them due to the capacity of the image file. The p-value for testing that the model has a correct functional form in *wbc* is 0.055, whereas the p-value corresponding to right model specification in terms of *crp* is given by 0.108. The results suggest that there is a certain degree of model mis-specification for the proportional odds model with the covariates *wbc* and *crp* but not with the co-

Table 7.8: The p-values of testing model mis-specification based on graphical diagnostics for model "age+smk+wbc+crp"

Tests	*age*	*smk*	*wbc*	*crp*	α (Bonferroni adjustment)
B_1	0.139	0.864	0.056	0.096	0.05 (0.025)
B_2	0.145	0.191	0.838	0.643	0.05 (0.025)
$(\mathbf{W}^m)_1$	0.175	0.981	0.055	0.108	0.05 (0.025)
$(\mathbf{W}^m)_2$	0.545	0.766	0.133	0.298	0.05 (0.025)
$(\mathbf{W}^*)_1$	0.175	0.981	0.055	0.108	0.05 (0.025)
$(\mathbf{W}^*)_2$	0.352	0.735	0.821	0.791	0.05 (0.025)
sum$(\mathbf{W}^m)$	0.352	0.735	0.821	0.791	0.05
max$(\mathbf{W}^m)$	0.299	0.799	0.069	0.188	0.05
prod$(\mathbf{W}^m)$	0.233	0.898	0.047	0.122	0.05
sum$(\mathbf{W}^*)$	0.332	0.866	0.193	0.209	0.05
max$(\mathbf{W}^*)$	0.235	0.829	0.059	0.156	0.05
prod$(\mathbf{W}^*)$	0.304	0.887	0.323	0.308	0.05

variates *age* and *smoking*. The raw scatter plots of actual FBG measurements on a continuous scale not included in the text also indicated a non-linear relationship between FBG and *wbc* and *crp*. Since the correlation between *wbc* and *crp* in the original dataset was very weak (0.10), we treat the model specification issue in each predictor separately, which may not be optimal in every situation. We may rather use $\mathbf{W}_o$ or a process containing only a few but not all covariates simultaneously. We discuss joint multivariate extensions of the proposed method in our concluding discussion.

As an illustration, we re-fit the proportional odds model including a quadratic and cubic term of *wbc* and a quadratic term in *crp* in Table 7.7. The linear and quadratic terms are significant in *wbc* with the cubic term marginally significant. The linear term in *crp* is also significant in the new model. The results corresponding to *age* and *smoking* remain almost unchanged in the second model, with both being non-significant. Table 7.9 presents the p-value of each of the graphical diagnostics for the model including higher order powers of *wbc* and *crp*. The graphic

Figure 7.4: Plot of residuals against *wbc* using the method $(\mathbf{W}^m)_1$ to check the model mis-specification for *wbc* in the model of "age+smk+wbc+crp". The dark black line indicates the observed process and the fine lines indicate the simulated realisations.

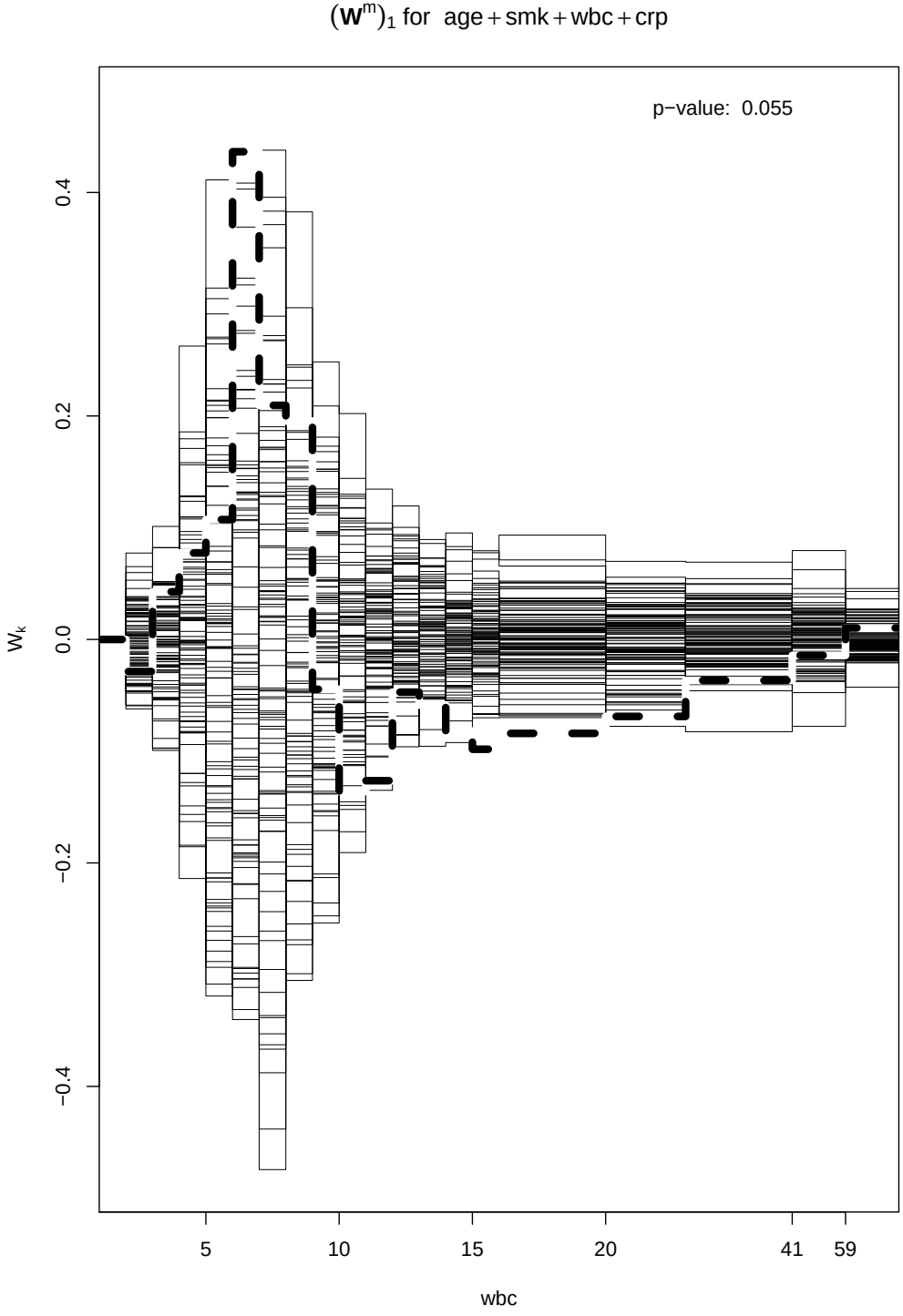


Figure 7.5: Plot of residuals against *crp* using the method $(\mathbf{W}^m)_1$ to check the model mis-specification for *crp* in the model of "age+smk+wbc+crp". The dark black line indicates the observed process and the fine lines indicate the simulated realisations.

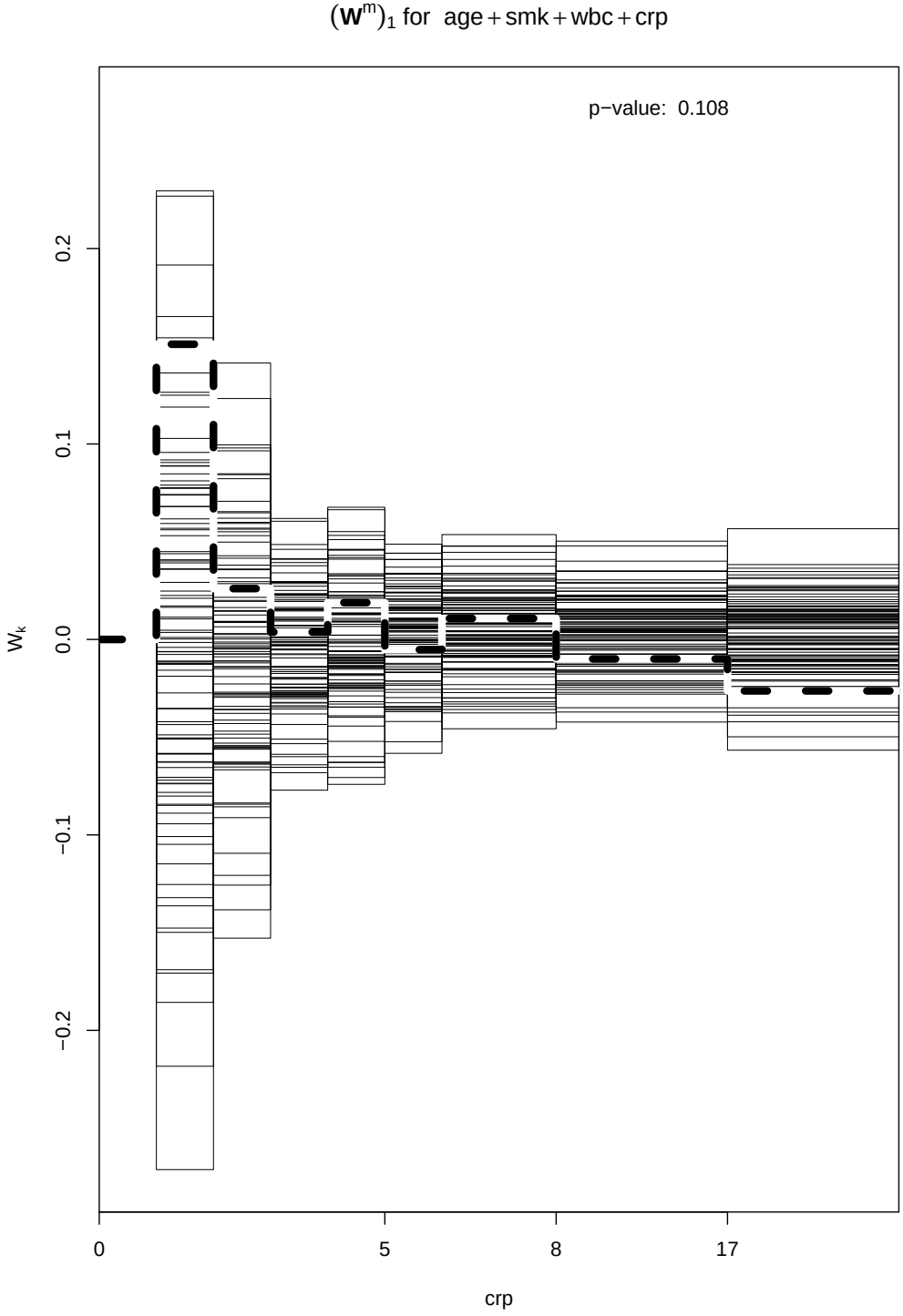


Table 7.9: The p-values of testing model mis-specification based on graphical diagnostics for model "age+smk+wbc+wbc^2+wbc^3+crp+crp^2"

Tests	*age*	*smk*	*wbc*	*crp*	α (Bonferroni adjustment)
B_1	0.262	0.774	0.497	0.662	0.05 (0.025)
B_2	0.249	0.148	0.125	0.071	0.05 (0.025)
$(\mathbf{W}^m)_1$	0.114	0.961	0.510	0.761	0.05 (0.025)
$(\mathbf{W}^m)_2$	0.543	0.875	0.532	0.678	0.05 (0.025)
$(\mathbf{W}^*)_1$	0.114	0.961	0.510	0.760	0.05 (0.025)
$(\mathbf{W}^*)_2$	0.334	0.712	0.347	0.231	0.05 (0.025)
sum$(\mathbf{W}^m)$	0.334	0.712	0.347	0.231	0.05
max$(\mathbf{W}^m)$	0.235	0.914	0.696	0.811	0.05
prod$(\mathbf{W}^m)$	0.196	0.943	0.679	0.699	0.05
sum$(\mathbf{W}^*)$	0.344	0.745	0.255	0.267	0.05
max$(\mathbf{W}^*)$	0.169	0.818	0.581	0.376	0.05
prod$(\mathbf{W}^*)$	0.428	0.940	0.225	0.299	0.05

diagnostics do not show model mis-specification for the new model. Figure 7.6 shows the plot using the method $(\mathbf{W}^m)_1$ for the new model for *wbc*, and Figure 7.7 shows the same for *crp*. The p-values are 0.51 and 0.761, respectively, indicating that the functional terms chosen in the final model are satisfactory. In terms of the actual FBG data on a continuous scale, it appears that there is a positive association between FBG and *crp* and *wbc* values for lower values of *crp* and *wbc*, below a certain threshold, but the relationship actually reverses or becomes less pronounced for higher extreme levels of these biomarkers, thus overall showing a non-linear pattern. There appears to be a non-linear threshold effect in the association between FBP with both *crp* and *wbc* when we analysed the continuous FBG data as well (Table 7.7).

Figure 7.6: Plot of residuals against *wbc* using the method $(\mathbf{W}^m)_1$ to check the model mis-specification for *wbc* in the model of "age+smk+wbc+wbc^2+wbc^3+crp+crp^2". The dark black line indicates the observed process and the fine lines indicate the simulated realisations.

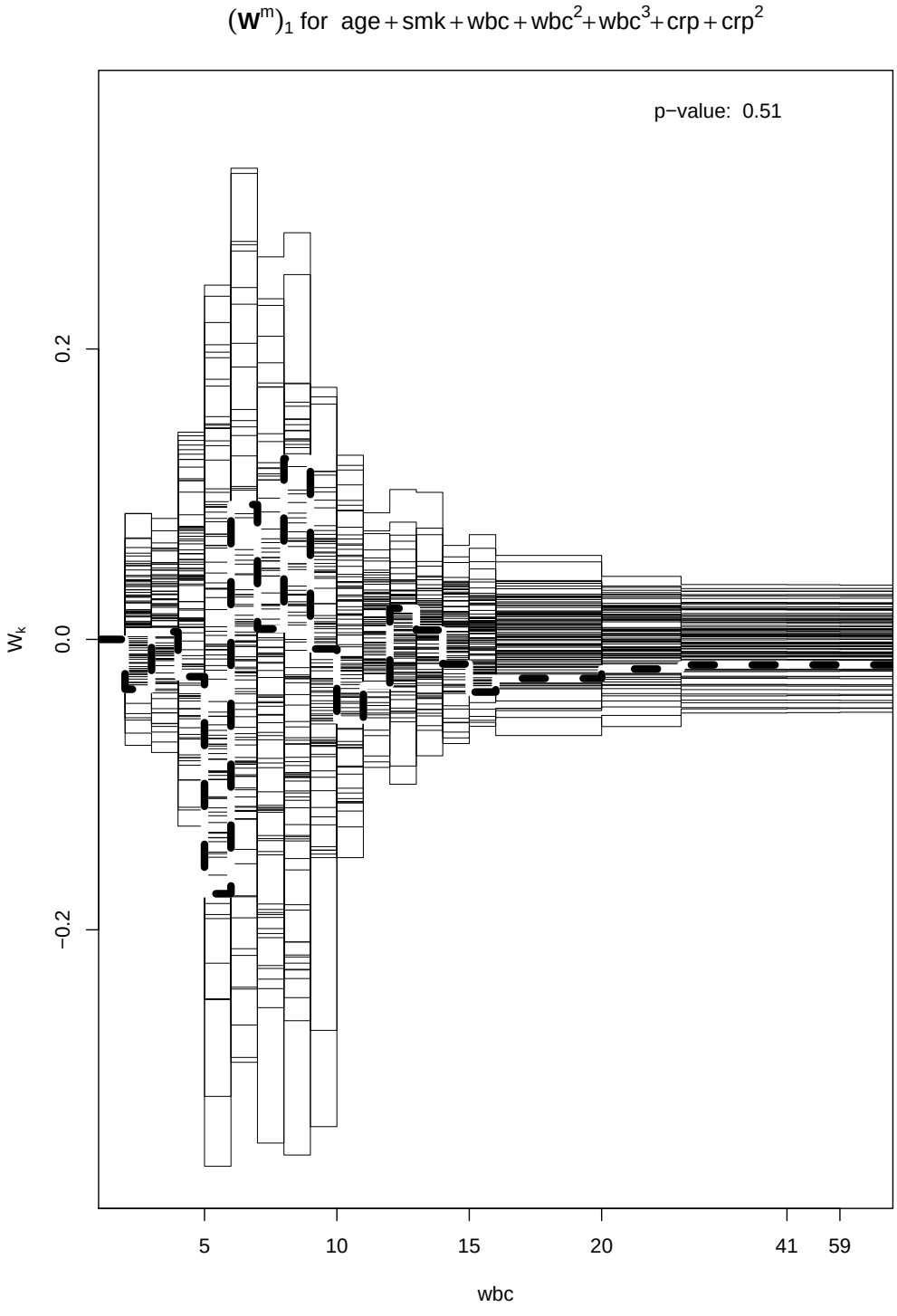
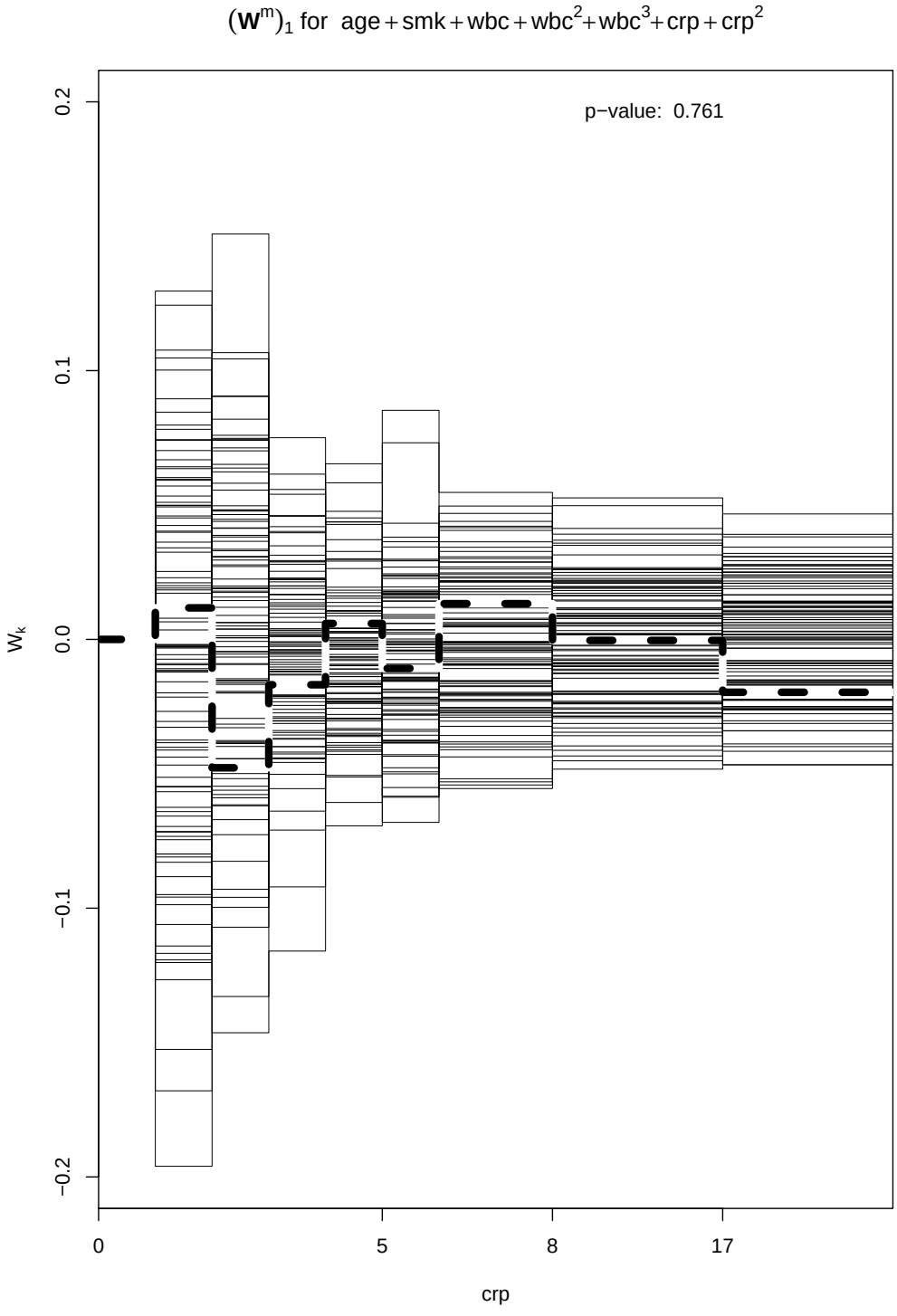


Figure 7.7: Plot of residuals against *crp* using the method $(\mathbf{W}^m)_1$ to check the model mis-specification for *crp* in the model of "age+smk+wbc+wbc^2+wbc^3+crp+crp^2". The dark black line indicates the observed process and the fine lines indicate the simulated realisations.



7.6 Discussion

Summary This chapter proposes graphical diagnostic methods based on two approaches to test model mis-specification for the proportional odds regression models. In the naive binary approach, we treat the proportional odds model as $J-1$ collapsed logistic regression models. Using the cumulative sums of residuals, the graphical diagnostic method extends previously introduced techniques by Lin et al. (2002) and Arbogast and Lin (2005). However, according to the simulations, it is more appropriate to treat the residuals in a multivariate format as in the second approach and then consider a vector of stochastic processes to represent the limiting behaviour of the residuals. In this manner, the asymptotic Gaussian processes ($\widehat{\mathbf{W}}_k$) take the correlation between the ordinal responses into account which is ignored in the binary approach.

In the multivariate approach, both the multivariate residuals ($\mathbf{r}$) and the cumulative residuals ($\mathbf{r}^*$) perform better than the binary approach but cumulative residuals outperform the multivariate residuals in our simulation study. For instance, in both scenarios, the methods based on $\mathbf{r}^*$ are better than the ones based on $\mathbf{r}$. Furthermore, among the different choices for the function to combine the components of a vector, $h(\cdot)$, the "sum" tends to be the best in most of our simulations.

Lin et al. (2002) noted that the tests are slightly more powerful when the process has the form

$$W_k^{(j)}(t;\hat{\boldsymbol{\beta}}) = n^{-1/2}\sum_{i=1}^{n}\hat{r}_{ij}^{*}\mathbb{1}(t-b<x_{ik}\leq t),$$

where b covers the lower half-plane of the covariates. In the large number of our simulations that there is not space to report on, including b does not give

consistently higher power. In general, we suggest taking $b = \infty$.

For a broad range of applications, we can use $\widehat{\mathbf{W}}_k(t; \hat{\boldsymbol{\gamma}})$ to a general multivariate generalised linear model and then use a function to combine the components of the multivariate residuals (or processes). These methods provide a good alternative to check the model fit and whether the chosen functional term is satisfactory. Simulation studies indicate they have power advantages compared to standard Hosmer-Lemeshow type partition-based statistic.

To conclude, in clinical investigations, as in the NAS example, investigators are often misled about the true nature of association between a predictor and a response due to fitting an incorrect model. For categorical responses, the task is even more daunting as there is no clear mandate about a single goodness-of-fit statistic. These simple graphical tools may provide us better insight into the inadequacies of the fitted model in such situations. The pattern in these plots may suggest alternative functional terms to include.

Extensions We mainly focused on the process $\mathbf{W}_k$, checking the functional form of a covariate of the proportional odds model. For alternative multinomial logit models to analyse ordinal response data discussed by Liu and Agresti (2005), such as adjacent-categories logit models and continuation-ratio logit models, one can extend the multivariate approach to make the graphical diagnostics in a similar manner. How to extend these tools to correlated ordinal responses is an interesting avenue for possible research (Pan 2002).

Sometimes, as in the Normative Aging Study, two covariates seem to be misspecified when considering only main effects. Instead of focusing on a single covariate, it seems wiser to focus jointly on two covariates by considering the

process

$$\mathbf{W}_o^m(t;\hat{\boldsymbol{\gamma}}) = n^{-1/2}\sum_{i=1}^{n}\mathbb{I}([x_{ik_1}, x_{ik_2}] \leq \mathbf{t})\hat{\mathbf{r}}_i,$$

where k_1 and k_2 are those covariates. Generally, any number of covariates $\leq p$ can be included. Similar to process $\mathbf{W}_k$, these processes are special cases of the process $\mathbf{W}_o$ and do not need further proofs. However, it remains to show how effective these new processes are.

Multiple Response Data Chapters 5 and 6 focused on the modelling of multiple response data. For each item j, we assumed a different marginal model of the form

$$g_j(\mu_{ij}) = \mathbf{z}_{ij}^T\boldsymbol{\beta}_j,\ j = 1, \ldots, J.$$

The most appealing fitting approach is GEE. The current chapter developed multivariate graphical diagnostic methods for GEE (and not only for ordinal data) which can also be applied to (repeated) multiple response data. In our view, it seems wise to consider a J-dimensional cumulative residual process, where the jth component refers to the jth model. In this way, we can check the mis-specification of J models simultaneously. If one would apply the univariate approach suggested by Lin et al. (2002), then no information is provided concerning which of the J marginal models is eventually mis-specified. Significance of the test would lead to an unsatisfactory rejection of all J models, although the majority of the J models might be correctly specified. As an alternative naive approach, one could apply the cumulative residual process for each of the J models separately, however, as with parameter estimation, the simultaneous approach accounts for dependence between items and is expected to have better properties than the naive approach.

Chapter 8

Conclusion

8.1 Odds Ratio Estimation

For stratified multiple response data, we considered three ways of defining the common odds ratio, a summarising measure for the conditional association between a row variable and the multiple response variable, given a stratification variable.

Greenland (1989) considered a generalised MH estimator by averaging over ordinary Mantel Haenszel (MH) estimators following the Mickey and Elashoff (1985) approach for estimating the common log odds ratio. He considered two sampling situations: One assumes (a) J independent multinomials per stratum, and another (b) J independent binomials per stratum, both forming K $2 \times J$ tables for which the MH estimators are dually consistent, consistent under limiting model I (large stratum sample size, while number of strata K is fixed) and limiting model II (where K becomes large, while sample size within strata is fixed). Greenland also derived (co)variance estimators for the ordinary and the generalised MH estimator that are dually consistent and valid for sampling models (a)

and (b).

In Chapter 2, we considered for each item $x = 1, \ldots, J$, the K $2 \times r$ tables formed by the positive and negative responses for each of the r rows and K strata. In such a way, we obtain r independent binomials per stratum for item x defining the kth odds ratio $\Psi^x_{ab|k}$ in terms of one item and two rows, item x and rows a and b. However, for two items x and y, the MH estimators $\hat{\Psi}^x_{ab}$ and $\hat{\Psi}^y_{ac}$ are not independent, and we derived a new dually consistent covariance estimator for the covariance between $\hat{\Psi}^x_{ab}$ and $\hat{\Psi}^y_{ac}$.

Another approach, called the model-based approach, treats the J items as a J-dimensional binary response vector and then uses logit models directly for the marginal distribution of each item. The parameter estimates can also be used as estimators for the common odds ratio. For model fitting, we applied the methodology of generalised estimation equations (GEE), a multivariate extension of the quasi-likelihood method, to account for dependency between items. The MH type estimators can also be considered as a non-model-based approach, because they estimate the odds ratios directly.

We investigated the performance of the MH-type estimators, the bootstrap estimators of (co)variance and the model-based estimators for a variety of configurations under independence and dependence of strata. The results confirm the good properties of the various MH estimators. Only under high dependence of strata, the bootstrap estimator and the model-based estimators outperform the MH estimators, which was expected, because the MH estimators are derived under the assumption of independence between strata.

In Chapter 3, we extended case (a) to (a'): Two independent rows of multiple response data per stratum, defining the kth odds ratio $\Psi_{xy|12k}$ in terms of two rows (rows 1 and 2) and two columns, the same way Greenland did for situation

(a). We showed that the MH estimator $\hat{\Psi}_{xy}$ is still dually consistent under (a'), but Greenland's (co)variance estimators are no longer dually consistent. Then we derived new dually consistent (co)variance estimators that are a generalisation of Greenland's (old) estimators, because his (co)variance estimators are special cases of ours. A simulation study confirms that the new (co)variance estimators are superior to the old estimators. Only when sampling under (a), the old and new estimators are identical and have equal performances, otherwise the new estimators perform much more strongly than the old. Unless sample sizes are very small (K and N_k), the new estimators also perform better than the bootstrap estimators of (co)variance. Unfortunately, we we are not aware of any model-based estimators, and could not compare the MH and its new (co)variance estimators with such a model-based approach.

Chapter 4 considers case (b'): One row of multiple response data per stratum, which can be considered as J dependent binomials, an extension of case (b). The ordinary MH estimator can still be applied but is only consistent under limiting model I. We proposed a new dually consistent MH estimator for estimating the common odds ratio. For this estimator, we also derived a dually consistent variance estimator. Due to the complex calculations, we decided to propose only a variance estimator with no covariance estimator. The variance estimator has a simple form, but each term of the asymptotic variance was estimated by only one term and not by averaging over several terms, as Greenland's (co)variance estimators and those in Chapter 3 were constructed, yielding a variance estimator that is less efficient.

The simulation study showed that the new MH estimator performs much better than the ordinary MH estimator except under independence. The new variance estimator also performs better than Greenland's estimator and the bootstrap

estimator of variance, except when tables are sparse and K is small. In those cases the bootstrap estimator of variance performs better. When items are indeed independent Greenland's estimator performs better. For situation (b'), there is also a model-approach using a logit model to estimate the odds ratio. However, the log odds ratio estimator performs badly and its true variance is significantly larger than that of the new MH estimator. Therefore the model-based approach cannot be recommended for this sampling situation. The generalised MH estimator can also be constructed from averaging over the newly proposed MH estimators, however, we cannot estimate the (co)variance of the generalised MH estimator, because we lack estimators for the covariance of two (new) MH estimators.

The odds ratios of Chapters 2 and 4 are defined in terms of positive and negative probabilities. The practitioner should be aware that the (co)variance estimators presented there are invariant under exchanging positive with negative responses. However the local odds ratio defined in Chapter 3 is defined in terms of positive probabilities only, similar to the relative risk. Therefore a subject-matter researcher must be aware of the meaning when exchanging positive with negative responses and applying any of the MH type estimators presented in Chapter 3.

8.2 HLP Diagnostics

As in the model-based approach, we can treat the J items as a J-dimensional binary response vector and then directly model the marginal distribution of each item in terms of some explanatory variables. Since we model the means of the univariate marginal distributions of the underlying multiple response variable, this type of modelling is also called marginal modelling. Modelling strategies

such as generalised linear models (GLM) can be applied to each of the J items.

In Chapter 5, we investigated deletion diagnostics for the marginal modelling approach expressing the marginal model as a homogeneous linear predictor (HLP) model, which is based on maximum likelihood (ML) estimation. The marginal model can also be fitted by generalised estimating equations (GEE) yielding more efficient estimates than fitting a GLM, which naively assumes independence between items.

For GEE the link function applies to the mean responses of the items and is one to one. For HLP, the link function maps from the expected counts of the joint table to the linear predictor and is many-to-one. Multiple case deletion for HLP models is different from GEE. We mainly focused on the Cook distance as a measure of influence. For HLP models and deletion of predictors, we considered three equivalent methods and concluded that the "delete=augment" method is our preferred method, because only the design matrix needs to be manipulated according to the deleted predictors.

For deletion of joint observations, we considered a standardised Cook distance, dividing the Cook distance by the number of multiple responses being deleted to account for those observations that are recorded multiple times.

8.3 Modelling of Repeated Multiple Response Data

The modelling of a repeated multiple response variable was considered in Chapter 6, where we distinguished between the marginal model approach and the random effect model approach. Unfortunately, ML methodology as HLP are not applicable anymore for the marginal model approach due to the large number of parameters describing the underlying joint distribution. In contrast, the GEE

method is still easily applicable. We considered several possible working correlation structures for the GEE approach to account for dependence between items and between occasions, and proposed a groupwise method, a simple correlation model assuming different groups have different values of correlation. If this assumption is true, the groupwise method has efficiency advantages over the standard method, which naively assumes that the correlation structure is equal for all subjects. The random effect approach is an alternative, but it can only incorporate non-negative correlations which might lead to inaccurate results if some correlation parameters of the data are negative. We illustrated the method using the STAT 291 data, a survey among students of the statistics lecture STAT 291 about their favourite bars, recording responses to questions related to age, sex, possible reasons for going out, favourite music, etc. Fitting the various models also showed that the groupwise and standard methods give substantially different results in terms of significance when groups are determined by the variable sex.

8.4 Graphical Diagnostic Method for Proportional Odds Model

In Chapter 7, we proposed two different approaches to investigate the mis-specification of a specific covariate for the proportional odds model. The binary approach considers the proportional odds model as $J-1$ logistic regression models and applies the cumulative residual process $W_k^{(j)}(t;\boldsymbol{\beta}_j)$ introduced by Arbogast and Lin (2005) for logistic regression to each of the $J-1$ logistic models. For each collapsed response j, large values of the supremum statistic G_{W_k} indicate such a mis-specification. A p-value can be obtained by computing the proportion of

the simulated realisations from a second process $\widehat{W}_k^{(j)}$, which is asymptotically equivalent to $W_k^{(j)}$, for which G_{W_k} exceeds $G_{\widehat{W}_k}$. To see a better picture of the mis-specification, we can also plot the residual process $W_k^{(j)}$ along with an artificial sample from a second process $\widehat{W}_k^{(j)}$ versus the kth covariate. If the cumulative residual process is relatively large in absolute value, then there is an indication of a mis-specified functional form of the kth covariate. We applied the Bonferroni method to adjust for the significance level while combining inference from all these plots.

In the multivariate approach, the proportional odds model is viewed as a member of the class of multivariate generalised linear models (MGLM), where the response variable is a vector of indicator responses. Consequently, the residual defined as the difference of mean responses and the vector of indicator responses is a $J-1$ dimensional vector. We considered a multivariate cumulative residual process $\mathbf{W}_k$ consisting of those multivariate residuals to assess the mis-specification of a specific covariate. Since the process is now multivariate, there are several ways of obtaining a p-value. One option is to obtain $J-1$ p-values by considering each component of the process separately. Then again the Bonferroni method can be applied. In such a way, the method also gives $J-1$ plots as in the binary approach. A better option is to consider the supremum statistic $G_{\mathbf{W}_o} = \sup_{\mathbf{t}\in\mathbb{R}^p} \|\mathbf{W}_o(\mathbf{t};\mathbf{b},\hat{\boldsymbol{\beta}})\|$ based on any norm $\|\cdot\|$. Such a norm plots the $J-1$ dimensional values of the multivariate cumulative residual process to the real plane. Generally, we can apply a function $h : \mathbb{R}^{J-1} \to \mathbb{R}$ to the multivariate residual process $\mathbf{W}_k$ to yield a univariate process. Then a single p-value can be easily obtained by considering a supremum statistic of $h(\mathbf{W}_k)$ and comparing its value relative to those from another asymptotically equivalent cumulative process $h(\widehat{\mathbf{W}}_k)$. This method also provides a single plot assessing the mis-

specification of a specific covariate.

The simulation study showed that the processes sum($\mathbf{W}_k^*$) and prod($\mathbf{W}_k^*$) yielded best results, where $\mathbf{W}_k^*$ is the process based on the multivariate cumulative residuals $\mathbf{r}_i^*$. This can be expected, because the proportional odds model is expressed in terms of cumulative probabilities. The naive binary collapsing approach exhibits the worst performance. It fails to maintain the nominal Type I error level and the estimated Type I error rate is twice the desired level of significance.

The process sum($\mathbf{W}_k$) was already considered by Lin et al. (2002) for GEE. Therefore, our cumulative residual processes can be seen as extensions of their processes, because we prove results for GEE and not only for the proportional odds model. Although we focused mainly on the mis-specification of a specific covariate, we also proposed processes for checking the functional form of the link function and of the overall model adequacy.

The methods were illustrated on two examples and worked well. The methods first indicated that the functional form of some covariates were mis-specified and then after a modification suggested that the final chosen functional form of these covariates was satisfactory.

8.5 Future Work

Some of the research can be further extended. First we consider the odds ratio estimation for stratified multiple response data. Under sampling model (b'), the variance estimator of the new MH estimator can still be improved by estimating each term of the asymptotic variances by averaging over several terms, the way Greenland (1989) constructed his (co)variance estimators. This is in contrast to

the way we did construct the newly proposed (co)variance estimators in Chapter 3. Another goal is to find dually consistent (co)variance estimators for the generalised MH estimator. However, this requires the asymptotic covariances of the new MH estimator under models I and II. For model II, this is even more complex than for the asymptotic variance, because the covariances refer to three or four items, whereas the variance refers only to two. This means we have to consider the joint distribution of three and four items consisting of $2^3 = 8$ and $2^4 = 16$ probabilities. For two items, the joint distribution was determined by $2^2 = 4$ joint probabilities only.

Greenland's (co)variance estimators for the generalised MH estimators have the same form for sampling models (a) and (b), but our estimators do have different forms for (a') and (b'). Ideally, we would find dually consistent estimators that are applicable for cases (a') and (b'), simultaneously. Greenland (1989) also introduced a generalised MH estimator for the person-time rate ratio, a ratio of two probabilities, and a generalised (co)variance estimator for this generalised MH estimator under sampling models (a) and (b). In the same way as we considered the three types of odds ratios, we could also extend the estimation of rate ratios to multiple response data for situations (a') and (b').

The various variance estimators are used to construct confidence intervals of the Wald-type. The question arises now how the coverage of these intervals is and how the Wald-type intervals perform compared to other types, such as the Wilson-type (Wilson 1927), and intervals based on resampling methods, such as bootstrapping and permutations.

Now let us focus on possible future work for the graphical diagnostic method based on multivariate cumulative residual processes. We already discussed some possible future research on page 276. We mainly focused on the process $\mathbf{W}_k$

checking the functional form of a covariate of the proportional odds model. The first question is how these methods perform for alternative multinomial logit models, such as adjacent-categories logit models and continuation-ratio logit models, to analyse ordinal response data as discussed by Liu and Agresti (2005). How to extend these tools to correlated ordinal responses is also an interesting avenue for possible research (Pan 2002).

We focused on the processes $\mathbf{W}_k$, but also proposed the processes $\mathbf{W}_o$ and $\mathbf{W}_p$. The process $\mathbf{W}_k$ focuses on the mis-specification of the kth covariate, whereas $\mathbf{W}_o$ focuses on the overall model mis-specification or, more precisely, on the mis-specification of all covariates simultaneously. In the same way, we could also consider such a process focusing on two or more covariates only, but not on all covariates. For example, it would be interesting to know whether such a process focusing on two covariates jointly performs worse or better than two processes $\mathbf{W}_k$ also focusing on the same two covariates.

The process $\mathbf{W}_k$ is defined as the sum over those residuals for which the kth covariate is less than or equal to a certain value t_k. We could also replace "less than or equal to" with "greater than". The process would have a completely different form, but still be applicable. The first process represents one path, the second another path. In fact, we could also consider any path from summing the residuals in different ways. We cannot compute the supremum over all such paths, because, for large data sets and continuous covariates, the total number of such paths becomes too large. Instead we might consider a limited number or a random sample from all such paths to yield a more robust test.

In Chapters 5 and 6, we focused on the modelling of multiple response data. The joint model comprised J marginal models can be fitted with the GEE method. Hence the cumulative residual processes also apply for this marginal model.

There are several questions: How do these processes perform for multiple response data and the marginal modelling? Is simultaneous model checking for all J models better than checking the models separately? We expect the simultaneous approach to perform better, but the analysis seems more complex. Therefore, it is debatable which approach is to be recommended.

Another problem is that we do not know the exact distribution of the cumulative residual process. We must sample from another process to approximate its distribution. Although this resampling is computationally feasible, we would prefer to construct a process with a known distribution. A similar process with limiting distribution $N(0,1)$ was proposed by Khmaladze and Koul (2004). It needs to be investigated how the processes $\mathbf{W}_p$, $\mathbf{W}_k$ and $\mathbf{W}_o$ can be constructed to have the same limiting distribution.

Appendix A

Derivation of the Asymptotic Variance for Model I of MH estimator - Chapter 3

In this part of the appendix, we want to derive the asymptotic variance of the MH estimator $\hat{\Psi}_{xy|ab}$ under the "large-stratum" limiting model (model I) by applying the delta method, see Subsection 3.3.1 on page 95.

We have $N = \sum_k N_k$ and as $N \to \infty$ $N\alpha_{ak} = n_{ak}$, where $0 < \alpha_{ak} < 1$. It follows that $N_k = \sum_i n_{ik} = N \sum_i \alpha_{ik}$. We prove the general case with $r > 2$ rows.
The MH estimator has the following form

$$\hat{\Psi}_{xy|ab} = \frac{\sum_k X_{x|ak} X_{y|bk}/N_k}{\sum_k X_{y|ak} X_{x|bk}/N_k} = \frac{\sum_k \frac{n_{ak}n_{bk}}{NN_k}\left(\frac{X_{A|a}}{n_{ak}} + \frac{X_{C|a}}{n_{ak}}\right)\left(\frac{X_{B|bk}}{n_{bk}} + \frac{X_{C|bk}}{n_{bk}}\right)}{\sum_k \frac{n_{ak}n_{bk}}{NN_k}\left(\frac{X_{B|a}}{n_{ak}} + \frac{X_{C|a}}{n_{ak}}\right)\left(\frac{X_{A|bk}}{n_{bk}} + \frac{X_{C|bk}}{n_{bk}}\right)}$$

with $X_{A|ak} = X^{10}_{xy|ak}$, $X_{B|ak} = X^{01}_{xy|ak}$ and $X_{C|ak} = X^{01}_{xy|ak}$.

Let the sample proportions be defined as $p_{ak} = X_{ak}/n_{ak}$ and vector $\mathbf{p}$ as $\mathbf{p} = (\mathbf{p}_1^T, \ldots, \mathbf{p}_K^T)^T$ with $\mathbf{p}_k = (\mathbf{p}_{1k}^T, \ldots, \mathbf{p}_{rk}^T)^T$ and $\mathbf{p}_{ak} = (p_{A|ak}, p_{B|ak}, p_{C|ak})^T$, such

that $\mathbf{p}$ contains all sample proportions. Similarly define vector $\boldsymbol{\pi}$ containing all probabilities $\pi_{A|ak}$, $\pi_{B|ak}$, and $\pi_{C|ak}$.

We want to argue that $\sqrt{N} \cdot \hat{\Psi}_{xy|ab}$ and $\sqrt{N} \cdot g(\boldsymbol{p})$ with

$$g(\boldsymbol{p}) = \frac{\sum_k (\sum_i \alpha_{ik}^{-1})^{-1} \frac{X_{x|ak}}{n_{ak}} \frac{X_{y|bk}}{n_{bk}}}{\sum_k (\sum_i \alpha_{ik}^{-1})^{-1} \frac{X_{y|ak}}{n_{ak}} \frac{X_{x|bk}}{n_{bk}}} = \frac{\sum_k (\sum_i \alpha_{ik}^{-1})^{-1} \left(\frac{X_{A|a}}{n_{ak}} + \frac{X_{C|a}}{n_{ak}}\right) \left(\frac{X_{B|bk}}{n_{bk}} + \frac{X_{C|bk}}{n_{bk}}\right)}{\sum_k (\sum_i \alpha_{ik}^{-1})^{-1} \left(\frac{X_{B|a}}{n_{ak}} + \frac{X_{C|a}}{n_{ak}}\right) \left(\frac{X_{A|bk}}{n_{bk}} + \frac{X_{C|bk}}{n_{bk}}\right)}$$

have the same limiting distributions. The kth summands $\sqrt{N} \cdot g_k^{num} := \sqrt{N} \cdot (\sum_i \alpha_{ik}^{-1})^{-1} \frac{X_{y|ak}}{n_{ak}} \frac{X_{x|bk}}{n_{bk}}$ and $\hat{\Psi}_k^{num} := \sqrt{N} \cdot \frac{n_{ak} n_{bk}}{N N_k} \frac{X_{x|ak}}{n_{ak}} \frac{X_{y|bk}}{n_{bk}}$ of the numerators of $\sqrt{N} \cdot \hat{\Psi}$ and $\sqrt{N} \cdot g(\mathbf{p})$ have the same limiting distributions by Slutsky's theorem, because the factor $\frac{n_{ak} n_{bk}}{N N_k}$ converges to $\alpha_{ak}\alpha_{bk}/(\sum_{i=1}^r \alpha_{ik})$. For $r = 2$, $\alpha_{ak}\alpha_{bk}/(\sum_{i=a,b} \alpha_{ik}) = (\sum_{i=a,b} \alpha_{ik}^{-1})^{-1}$. Although we prove the general case $r > 2$, we write for convenience $(\sum_i \alpha_{ik}^{-1})^{-1}$ instead of $\alpha_{ak}\alpha_{bk}/(\sum_{i=1}^r \alpha_{ik})$, which is only a technical matter.

By the multivariate C.L.T. (Theorem 2.8.6 on page 75), the sample proportions from the multinomial distributions are asymptotically multivariate normally distributed

$$\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Sigma})$$

with $\boldsymbol{\Sigma} = \text{Diag}(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$, $\boldsymbol{\Sigma}_k = \text{Diag}(\frac{1}{\alpha_{1k}} \boldsymbol{\Sigma}_{1k}, \ldots, \frac{1}{\alpha_{rk}} \boldsymbol{\Sigma}_{rk})$ and $\boldsymbol{\Sigma}_{ak} = \text{Diag}(\boldsymbol{\pi}_{ak}) - \boldsymbol{\pi}_{ak}\boldsymbol{\pi}_{ak}^T$. It follows, $\sqrt{N} \cdot g_k^{num}$ and $\sqrt{N} \cdot \hat{\Psi}_k^{num}$ also converge to the same normal distribution. Now because this limiting distribution is normal for all k, the sums $\sqrt{N} \cdot \sum_k g_k^{num}$ and $\sqrt{N} \cdot \sum_k \hat{\Psi}_k^{num}$ also have the same limiting normal distribution. By noting that the denominators of $\hat{\Psi}$ and $g(\mathbf{p})$ converge to the same constant (the sample proportions $p_{ak} = X_{ak}/n_{ak}$ converge to π_{ak}), $\sqrt{N} \cdot \hat{\Psi}_{xy|ab}$ and $\sqrt{N} \cdot g(\boldsymbol{p})$ also have the same limiting normal distribution by Slutsky's theorem (Theorem 2.8.1 on page 74).

Estimator $\hat{\Psi}$ converges in probability to $g(\boldsymbol{\pi}) = \frac{\lim_N \mathbb{E} C_{xy|ab}/N}{\lim_N \mathbb{E} C_{yx|ab}/N}$, which equals Ψ

under the common odds ratio assumption, see proof of Theorem 3.2.1 on page 92 for details.

The delta method (Theorem 2.8.4) says $\sqrt{N}(g(\mathbf{p})-g(\boldsymbol{\pi})) \xrightarrow{d} N(\mathbf{0}, V_g = \frac{\partial g}{\partial \pi}\boldsymbol{\Sigma}\frac{\partial g}{\partial \pi}^T)$. In the following, we write shorter $\partial g/\partial \pi$, but mean $\partial g/\partial p|_{p=\pi}$.

The kth odds ratio is defined as

$$\Psi_k = \pi_{x|1k}\pi_{y|2k}/\pi_{y|1k}\pi_{x|2k} = A_k/B_k$$

with

$$A_k = \pi_{x|ak}\pi_{y|bk} = (\pi_{A|ak} + \pi_{C|ak})(\pi_{B|bk} + \pi_{C|bk}),$$

$$B_k = \pi_{y|ak}\pi_{xk|bk} = (\pi_{B|ak} + \pi_{C|ak})(\pi_{A|bk} + \pi_{C|bk})$$

Define $t_k := (\sum_i \alpha_{ik}^{-1})^{-1}B_k$ and $w_k := t_k/\sum_j t_j$. We express $g(\boldsymbol{\pi})$ as

$$g(\boldsymbol{\pi}) = \sum_{l=1}^{K} w_l\Psi_l = \sum_l \exp\left\{\log(\sum_i \alpha_{il}^{-1})^{-1} + \log(A_l) - \log[\sum_{j=1}^{K}(\sum_i \alpha_{ij}^{-1})^{-1}B_j]\right\}.$$

First we rewrite V_g as

$$\begin{aligned}
&V_g \\
&= \sum_{k=1}^{K}\sum_{i=1}^{2}\frac{1}{\alpha_{ik}}\{\frac{\partial g}{\partial \pi_{A|ik}}\Sigma_{ik}|_{A,A} + \frac{\partial g}{\partial \pi_{B|ik}}\Sigma_{ik}|_{B,B} + \frac{\partial g}{\partial \pi_{C|ik}}\Sigma_{ik}|_{C,C} \\
&+ \frac{\partial g}{\partial \pi_{A|ik}}\frac{\partial g}{\partial \pi_{B|ik}}\Sigma_{ik}|_{A,B} + \frac{\partial g}{\partial \pi_{A|ik}}\frac{\partial g}{\partial \pi_{C|ik}}\Sigma_{ik}|_{A,C} + \frac{\partial g}{\partial \pi_{B|ik}}\frac{\partial g}{\partial \pi_{C|ik}}\Sigma_{ik}|_{B,C}\} \\
&= \sum_{k=1}^{K}\sum_{i=1}^{2}\frac{1}{\alpha_{ik}}\left[\pi_{A|ik}\left(\frac{\partial g}{\partial \pi_{A|ik}}\right)^2 + \pi_{B|ik}\left(\frac{\partial g}{\partial \pi_{B|ik}}\right)^2 + \pi_{C|ik}\left(\frac{\partial g}{\partial \pi_{C|ik}}\right)^2\right] \\
&- \sum_{k=1}^{K}\sum_{i=1}^{2}\frac{1}{\alpha_{ik}}\left[\pi_{A|ik}\frac{\partial g}{\partial \pi_{A|ik}} + \pi_{B|ik}\frac{\partial g}{\partial \pi_{B|ik}} + \pi_{C|ik}\frac{\partial g}{\partial \pi_{C|ik}}\right]^2,
\end{aligned} \tag{A.1}$$

where $\Sigma_{ak}|_{m,n}$ denotes the mth row and nth column of Σ_{ak}.

Now we compute

$$\begin{array}{llll}
\frac{\partial A_k}{\partial \pi_{A|ak}} = \pi_{y|bk} & \frac{\partial A_k}{\partial \pi_{A|bk}} = 0 & \frac{\partial B_k}{\partial \pi_{A|ak}} = 0 & \frac{\partial B_k}{\partial \pi_{A|bk}} = \pi_{x|ak} \\
\frac{\partial A_k}{\partial \pi_{B|ak}} = 0 & \frac{\partial A_k}{\partial \pi_{B|bk}} = \pi_{x|ak} & \frac{\partial B_k}{\partial \pi_{B|ak}} = \pi_{x|bk} & \frac{\partial B_k}{\partial \pi_{B|bk}} = 0 \\
\frac{\partial A_k}{\partial \pi_{C|ak}} = \pi_{y|bk} & \frac{\partial A_k}{\partial \pi_{C|bk}} = \pi_{x|ak} & \frac{\partial B_k}{\partial \pi_{C|ak}} = \pi_{x|bk} & \frac{\partial B_k}{\partial \pi_{C|bk}} = \pi_{y|ak},
\end{array}$$

therefore, we have

$$\begin{array}{ll}
\frac{\partial g}{\partial \pi_{B|ak}} = -\sum_l w_l \Psi_l \frac{1}{\sum t_j} (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{x|bk} = & -\pi_{x|bk} \frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j} g(\boldsymbol{\pi}) \\
\frac{\partial g}{\partial \pi_{A|ak}} = \frac{w_k \Psi_k}{A_k} \pi_{y|bk} & \\
\frac{\partial g}{\partial \pi_{C|ak}} = \frac{w_k \Psi_k}{A_k} \pi_{y|bk} - \pi_{x|bk} \frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j} g(\boldsymbol{\pi}) = & \frac{\partial g}{\partial \pi_{ak}^A} + \frac{\partial g}{\partial \pi_{B|ak}} \\
\frac{\partial g}{\partial \pi_{B|bk}} = \frac{w_k \Psi_k}{A_k} \pi_{x|ak} & \\
\frac{\partial g}{\partial \pi_{A|bk}} = -\sum_l w_l \Psi_l \frac{1}{\sum t_j} (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{y|ak} = & -\pi_{y|ak} \frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j} g(\boldsymbol{\pi}) \\
\frac{\partial g}{\partial \pi_{C|bk}} = \frac{w_k \Psi_k}{A_k} \pi_{x|ak} - \pi_{y|ak} \frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j} g(\boldsymbol{\pi}) = & \frac{\partial g}{\partial \pi_{A|bk}} + \frac{\partial g}{\partial \pi_{B|bk}}.
\end{array}$$

Now

$$\begin{aligned}
&\pi_{A|a} \frac{\partial g}{\partial \pi_{A|ak}} + \pi_{B|ak} \frac{\partial g}{\partial \pi_{B|a}} + \pi_{C|a} \frac{\partial g}{\partial \pi_{C|a}} = \\
&= \pi_{A|a} \frac{w_k \Psi_k}{A_k} \pi_{y|bk} - \pi_{B|a} \pi_{x|bk} \frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j} g(\boldsymbol{\pi}) \\
&+ \pi_{C|a} \frac{w_k \Psi_k}{A_k} \pi_{y|bk} - \pi_{C|a} \pi_{x|bk} \frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j} g(\boldsymbol{\pi}) \\
&= \pi_{x|ak} \pi_{y|bk} \frac{w_k \Psi_k}{A_k} - \frac{g(\boldsymbol{\pi})}{\sum t_j} (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{y|ak} \pi_{x|bk} \\
&= \frac{t_k}{\sum t_j} \Psi_k - \frac{t_k}{\sum t_j} g(\boldsymbol{\pi}) = w_k [\Psi_k - g(\boldsymbol{\pi})]. \qquad \text{(A.2)}
\end{aligned}$$

Similarly

$$\pi_{A|b}\frac{\partial g}{\partial \pi_{A|b}} + \pi_{B|b}\frac{\partial g}{\partial \pi_{B|b}} + \pi_{C|b}\frac{\partial g}{\partial \pi_{C|b}} = w_k[\Psi_k - g(\boldsymbol{\pi})]. \tag{A.3}$$

Now we simplify (A.1) by using (A.2) and (A.3)

$$\begin{aligned}
&V_g \\
&= \sum_{k=1}^{K}\sum_{i=a,b}\frac{1}{\alpha_{ik}}\left[\pi_{A|i}\left(\frac{\partial g}{\partial \pi_{A|i}}\right)^2 + \pi_{B|i}\left(\frac{\partial g}{\partial \pi_{B|i}}\right)^2 + \pi_{C|i}\left(\frac{\partial g}{\partial \pi_{C|i}}\right)^2\right] \\
&- \sum_{k=1}^{K}\sum_{i=a,b}\frac{1}{\alpha_{ik}}\left[\pi_{A|i}\frac{\partial g}{\partial \pi_{A|i}} + \pi_{B|i}\frac{\partial g}{\partial \pi_{B|i}} + \pi_{C|i}\frac{\partial g}{\partial \pi_{C|i}}\right]^2 \\
&= \sum_k \frac{1}{\alpha_{1k}}\pi_{B|a}(-\pi_{x|bk}\frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j}g(\boldsymbol{\pi}))^2 + \pi_{A|a}(\frac{w_k\Psi_k}{A_k}\pi_{y|bk})^2 \\
&+ \pi_{C|a}(\frac{w_k\Psi_k}{A_k}\pi_{y|bk} - \pi_{x|bk}\frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j}g(\boldsymbol{\pi}))^2 + w_k(\Psi_k - g(\boldsymbol{\pi}))\} \\
&+ \sum_k \frac{1}{\alpha_{2k}}\{\pi_{B|b}(\frac{w_k\Psi_k}{A_k}\pi_{x|ak})^2 + \pi_{A|b}(-\pi_{y|ak}\frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j}g(\boldsymbol{\pi}))^2 \\
&+ \pi_{C|b}(\frac{w_k\Psi_k}{A_k}\pi_{x|ak} - \pi_{y|ak}\frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j}g(\boldsymbol{\pi}))^2 + [w_k(\Psi_k - g(\boldsymbol{\pi}))]^2\} \\
&= \sum_k \frac{1}{\alpha_{1k}}\{\pi_{B|a}\pi_{x|bk}^2\frac{(\sum_i \alpha_{ik}^{-1})^{-2}}{(\sum t_j)^2}g(\boldsymbol{\pi})^2 + \pi_{A|a}\frac{w_k^2\Psi_k^2}{A_k^2}\pi_{y|bk}^2 + [w_k(\Psi_k - g(\boldsymbol{\pi}))]^2 \\
&+ \pi_{C|a}\frac{w_k^2\Psi_k^2}{A_k^2}\pi_{y|bk}^2 - 2\pi_{C|a}\frac{w_k\Psi_k}{A_k}\pi_{y|bk}\pi_{x|bk}\frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j}g(\boldsymbol{\pi}) + \pi_{C|a}\frac{(\sum_i \alpha_{ik}^{-1})^{-2}}{(\sum t_j)^2}g(\boldsymbol{\pi})^2\} \\
&+ \sum_k \frac{1}{\alpha_{2k}}\{\pi_{B|b}\frac{w_k^2\Psi_k^2}{A_k^2}\pi_{x|ak}^2 + \pi_{A|b}\pi_{y|ak}^2\frac{(\sum_i \alpha_{ik}^{-1})^{-2}}{(\sum t_j)^2}g(\boldsymbol{\pi})^2 + [w_k(\Psi_k - g(\boldsymbol{\pi}))]^2 \\
&+ \pi_{C|b}\frac{w_k^2\Psi_k^2}{A_k^2}\pi_{x|ak}^2 - 2\pi_{C|b}\frac{w_k\Psi_k}{A_k}\pi_{x|ak}\pi_{y|ak}\frac{(\sum_i \alpha_{1k}^{-1})^{-1}}{\sum t_j}g(\boldsymbol{\pi}) + \pi_{C|b}\pi_{y|ak}^2\frac{(\sum_i \alpha_{ik}^{-1})^{-2}}{(\sum t_j)^2}g(\boldsymbol{\pi})^2\} \\
&= \sum_k \frac{1}{\alpha_{1k}}\{\pi_{y|ak}\pi_{x|bk}^2\frac{(\sum_i \alpha_{ik}^{-1})^{-2}}{(\sum t_j)^2}g(\boldsymbol{\pi})^2 + \pi_{x|ak}\frac{w_k^2\Psi_k^2}{A_k^2}\pi_{y|bk}^2 \\
&+ \sum_k \frac{1}{\alpha_{2k}}\{\pi_{y|bk}\frac{w_k^2\Psi_k^2}{A_k^2}\pi_{x|ak}^2 + \pi_{x|bk}\pi_{y|ak}^2\frac{(\sum_i \alpha_{1k}^{-1})^{-2}}{(\sum t_j)^2}g(\boldsymbol{\pi})^2\} \\
&+ \sum_k(\sum_i \alpha_{ik}^{-1})[w_k(\Psi_k - g(\boldsymbol{\pi}))]^2
\end{aligned}$$

$$
\begin{aligned}
&- \sum_k \frac{2}{\alpha_{1k}} \{\pi_{C|b} \frac{w_k \Psi_k}{A_k} \pi_{x|ak} \pi_{y|ak} \frac{(\sum_i \alpha_{ik}^{-1})^{-1}}{\sum t_j} g(\boldsymbol{\pi})\} \\
&- \sum_k \frac{2}{\alpha_{bk}} \{\pi_{C|a} \frac{w_k \Psi_k}{A_k} \pi_{y|bk} \pi_{x|bk} \frac{(\sum_i \alpha_{1k}^{-1})^{-1}}{\sum t_k} g(\boldsymbol{\pi})\} \\
&= \frac{1}{(\sum t_k)^2} \sum_k \frac{1}{\alpha_{1k}} \{\pi_{y|ak} \pi_{x|bk}^2 \frac{t_k^2}{B_k^2} g(\boldsymbol{\pi})^2 + \pi_{x|ak} \pi_{y|bk}^2 \frac{t_k^2 w_k^2 \Psi_k^2}{A_k^2}\} \\
&+ \frac{1}{(\sum t_k)^2} \sum_k \frac{1}{\alpha_{2k}} \{\pi_{y|bk} \pi_{x|ak}^2 \frac{t_k^2 \Psi_k^2}{A_k^2} + \pi_{x|bk} \pi_{y|ak}^2 \frac{t_k^2}{B_k^2} g(\boldsymbol{\pi})^2\} \\
&+ \sum_k (\sum_i \alpha_{ik}^{-1}) [w_k (\Psi_k - g(\boldsymbol{\pi}))]^2 - \frac{2}{(\sum t_k)^2} \sum_k \frac{\pi_{C|a}}{\alpha_{1k}} \{\pi_{y|bk} \pi_{x|bk} \frac{t_k^2}{B_k A_k} \Psi_k g(\boldsymbol{\pi})\} \\
&- \frac{2}{(\sum t_k)^2} \sum_k \frac{\pi_{C|b}}{\alpha_{2k}} \{\pi_{x|ak} \pi_{y|ak} \frac{t_k^2}{A_k B_k} \Psi_k g(\boldsymbol{\pi})\} \\
=& \frac{1}{(\sum t_k)^2} \{\sum_k \frac{1}{\alpha_{1k}} [\pi_{y|bk} \frac{t_k^2 \Psi_k^2}{A_k} + \pi_{x|bk} \frac{t_k^2 g(\boldsymbol{\pi})^2}{B_k} - 2\pi_{C|a} \pi_{y|bk} \pi_{x|bk} \frac{t_k^2 \Psi_k g(\boldsymbol{\pi})}{A_k B_k}] \\
&+ \sum_k \frac{1}{\alpha_{2k}} [\pi_{x|ak} \frac{t_k^2 \Psi_k^2}{A_k} + \pi_{y|ak} \frac{t_k^2 g(\boldsymbol{\pi})^2}{B_k} - 2\pi_{C|b} \pi_{x|ak} \pi_{y|ak} \frac{t_k^2 \Psi_k g(\boldsymbol{\pi})}{A_k B_k}] \\
&- \sum_k (\sum_i \alpha_{1k}^{-1}) [w_k (\Psi_k - g(\boldsymbol{\pi}))]^2. \qquad \text{(A.4)}
\end{aligned}
$$

Under the common odds ratio assumption $\Psi = \Psi_1 = \cdots = \Psi_K$ and $g(\boldsymbol{\pi}) = \Psi$, consequently the term $[w_k(\Psi_k - g(\boldsymbol{\pi}))]^2 = [w_k(\Psi - \Psi)]^2 = 0$ and V_g can be written as

$$
\begin{aligned}
V_g &= \frac{\Psi^2}{(\sum t_k)^2} \sum_k \frac{t_k^2}{\alpha_{1k}} [(\frac{\pi_{y|bk}}{A_k} + \frac{\pi_{x|bk}}{B_k}) - 2\frac{\pi_{C|a} \pi_{x|bk} \pi_{y|bk}}{A_k B_k}] \\
&+ \frac{\Psi^2}{(\sum t_k)^2} \sum_k \frac{t_k^2}{\alpha_{2k}} [(\frac{\pi_{x|ak}}{A_k} + \frac{\pi_{y|ak}}{B_k}) - 2\frac{\pi_{C|b} \pi_{x|ak} \pi_{y|ak}}{A_k B_k}] \\
&= \frac{\sum_k (\sum_i \alpha_{ik}^{-1})^{-2} \frac{1}{\alpha_{1k}} [\pi_{x|ak} \pi_{y|bk}^2 + \Psi^2 \pi_{y|ak} \pi_{x|bk}^2 - 2\Psi \pi_{C|a} \pi_{x|bk} \pi_{y|bk}]}{(\sum_k (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{y|ak} \pi_{x|bk})^2} \\
&+ \frac{\sum_k (\sum_i \alpha_{ik}^{-1})^{-2} \frac{1}{\alpha_{2k}} [\pi_{x|ak}^2 \pi_{y|bk} + \Psi^2 \pi_{y|ak}^2 \pi_{x|bk} - 2\Psi \pi_{C|b} \pi_{x|ak} \pi_{y|ak}]}{(\sum_k (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{y|ak} \pi_{x|bk})^2} \\
&= \frac{\sum_k \frac{(\sum_i \alpha_{ik}^{-1})^{-2}}{\alpha_{1k}} [\pi_{x|ak} \pi_{y|bk}^2 + \Psi^2 \pi_{y|ak} \pi_{x|bk}^2 - 2\Psi \pi_{C|a} \pi_{x|bk} \pi_{y|bk}]}{(\sum_k (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{y|ak} \pi_{x|bk})^2}
\end{aligned}
$$

$$+\frac{\sum_k \frac{(\sum_i \alpha_{ik}^{-1})^{-2}}{\alpha_{2k}}[\pi_{x|ak}^2\pi_{y|bk}+\Psi^2\pi_{y|ak}^2\pi_{x|bk}-2\Psi\pi_{C|b}\pi_{x|ak}\pi_{y|ak}]}{(\sum_k(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{y|ak}\pi_{x|bk})^2}. \quad \text{(A.5)}$$

We conclude that under the common odds ratio assumption, $\sqrt{N}(\hat{\Psi}_{xy|ab}-\Psi_{xy|ab})$ is asymptotically Gaussian distributed with zero mean and variance $\lim_{N\to\infty} N \cdot \text{Var}^a(\hat{\Psi}_{xy|ab}) = V_g$.

Appendix B

Derivation of Multinomial/Binomial Distribution - Chapter 3

In this part of the Appendix, we want to show under which circumstances the multinomial and binomial distributions are special cases of the joint distribution of a multiple response (respectively any/J) variable. The following results are used at various stages of Chapter 3.

B.1 Multinomial Responses as Special Cases of Multiple Responses

First note, the multinomial distribution is a special case of the joint distribution of the multiple responses. For the multinomial distribution with J categories, we have the following joint probabilities

$$\Pr(Y_1 = 0, \ldots, Y_{x-1} = 0, Y_x = 1, Y_{x+1} = 0, \ldots, Y_J = 0|ak) = \pi_{x|ak}$$

with $\sum_{x=1}^{J} \pi_{x|ak} \leq 1$. Therefore, all remaining joint probabilities $\Pr(Y_1 = j_1, \ldots, Y_J = j_J|ak)$ with $j_k \in \{0, 1\}$ are zero, except $\Pr(Y_1 = 0, Y_2, \ldots, Y_J = 0|ak) > 0$ if $\sum_{x=1}^{J} \pi_{x|ak} < 1$. For example, we cannot observe the sequences $(1, 1, 1, \ldots, 1)$ and $(1, 1, 0, \ldots, 0)$. The covariance between two items (categories) of the multinomial distribution is $\text{Cov}(Y_x, Y_y) = -\pi_{x|ak}\pi_{y|ak}$. The binomial is also a special case of multiple responses, because the binomial is a special case of the multinomial.

B.2 Fixing the Covariance between Two Items

For two items x and y, we set $\text{Cov}(Y_x, Y_y) = -\pi_{x|ak}\pi_{y|ak}$, or in other words we set the covariance between two items so that it matches the covariance between two categories of a multinomial distribution.

From $-\pi_{x|ak}\pi_{y|ak} = \text{Cov}(Y_x, Y_y) = \mathbb{E}Y_xY_y - \mathbb{E}Y_x\mathbb{E}Y_y = \mathbb{E}Y_xY_y - \pi_{x|ak}\pi_{y|ak}$ follows $\mathbb{E}Y_xY_y = 0$. The variables Y_x are binary and by definition $\mathbb{E}Y_xY_y = \sum_{i,j=0}^{1} ij \Pr($ $Y_x = i, Y_y = j) = \Pr(Y_x = 1, Y_y = 1)$, and it follows that $\pi_{xy} = \Pr(Y_x = 1, Y_y = 1) = 0$. Therefore $\Pr(Y_x = 1, Y_y = 0|ak) = \pi_{x|ak} - \pi_{xy|ak} = \pi_{x|ak}$, $\Pr(Y_x = 0, Y_y = 1|ak) = \pi_{y|ak}$, and $\Pr(Y_x = 0, Y_y = 0|ak) = 1 - \pi_{x|ak} - \pi_{y|ak}$; see also (2.18) on page 73 for relations between the marginal and pairwise probabilities. For the special case $\pi_{x|ak} + \pi_{y|ak} = 1$, the response for item y is exactly the opposite of the response of item x, that is, responses for items x and y form a binary distribution. For $\pi_{x|ak} + \pi_{y|ak} < 1$, we yield a multinomial with 3 possible outcome categories.

B.3 Fixing the Covariance between More Than Two Items

Now we set for all pairs of items (*x*,*y*), with $x < y; x, y \in \{1, \ldots, J\}$, $\text{Cov}(Y_x, Y_y) = -\pi_{x|ak}\pi_{y|ak}$, which yields the multinomial distribution. We prove this by induction and use as a base case two items only.

Proof by Induction

Proposition

Let $(i_1, \ldots, i_m)$ be an arbitrary distinct index set of length m of the set $\{1, \ldots, J\}$. We omit indices i and k standing for the row and stratum.

$$\Pr(Y_{i_1} = 1, \ldots, Y_{i_l} = 1, Y_{i_{l+1}} = 0, \ldots, Y_{i_m} = 0) = 0 \text{ for } l \geq 2 \tag{B.1}$$

$$\Pr(Y_{i_1} = 1, Y_{i_2} = 0, \ldots, Y_{i_m} = 0) = \pi_x \tag{B.2}$$

$$\Pr(Y_{i_1} = 0, Y_{i_2} = 0, \ldots, Y_{i_m} = 0) = 1 - \sum_{j=1}^{m} \pi_{i_j} \tag{B.3}$$

Base Case

Under the condition $\text{Cov}(Y_x, Y_y) = -\pi_{x|ak}\pi_{y|ak}$, we derived the following pairwise probabilities

$$\Pr(Y_x = 1, Y_y = 1) = 0$$

$$\Pr(Y_x = 1, Y_y = 0) = \pi_x$$

$$\Pr(Y_x = 0, Y_y = 0) = 1 - \pi_x - \pi_y$$

for any two distinct indices $x, y \in \{1, \ldots, J\}$.

Inductive Hypothesis

$$\Pr(Y_{i_1} = 1, \ldots, Y_{i_l} = 1, Y_{i_{l+1}} = 0, \ldots, Y_{i_m} = 0, Y_{i_{m+1}} = 0) = 0 \text{ for } l \geq 2 \quad \text{(B.4)}$$

$$\Pr(Y_{i_1} = 1, Y_{i_2} = 0, \ldots, Y_{i_m} = 0, Y_{i_{m+1}} = 0) = \pi_x \quad \text{(B.5)}$$

$$\Pr(Y_{i_1} = 0, Y_{i_2} = 0, \ldots, Y_{i_m} = 0, Y_{i_{m+1}} = 0) = 1 - \sum_{j=1}^{m+1} \pi_{i_j} \quad \text{(B.6)}$$

for $m + 1 \leq J$.

Inductive Step

By (B.1)

$$\begin{aligned} 0 &= \Pr(Y_{i_1} = 1, \ldots, Y_{i_l} = 1, Y_{i_{l+1}} = 0, \ldots, Y_{i_m} = 0) \\ &= \Pr(Y_{i_1} = 1, \ldots, Y_{i_l} = 1, Y_{i_{l+1}} = 0, \ldots, Y_{i_m} = 0, Y_{i_{m+1}} = 0) \\ &+ \Pr(Y_{i_1} = 1, \ldots, Y_{i_l} = 1, Y_{i_{l+1}} = 0, \ldots, Y_{i_m} = 0, Y_{i_{m+1}} = 1) \geq 0, \end{aligned}$$

(B.4) follows. Now we can show (B.5)

$$\begin{aligned} &\Pr(Y_{i_1} = 1, Y_{i_2} = 0, \ldots, Y_{i_m} = 0, Y_{i_{m+1}} = 0) \\ &= \Pr(Y_{i_1} = 1, Y_{i_2} = 0, \ldots, Y_{i_m} = 0) - \Pr(Y_{i_1} = 1, Y_{i_2} = 0, \ldots, Y_{i_m} = 0, Y_{i_{m+1}} = 1) \\ &= \pi_x - 0 = \pi_x \end{aligned}$$

by (B.2) and (B.4). Finally we derive (B.6)

$$\begin{aligned} &\Pr(Y_{i_1} = 0, Y_{i_2} = 0, \ldots, Y_{i_m} = 0, Y_{i_{m+1}} = 0) \\ &= \Pr(Y_{i_1} = 0, Y_{i_2} = 0, \ldots, Y_{i_m} = 0) - \Pr(Y_{i_1} = 0, Y_{i_2} = 0, \ldots, Y_{i_m} = 0, Y_{i_{m+1}} = 1) \end{aligned}$$

$$= 1 - \sum_{j=1}^{m} \pi_{i_j} - \pi_{i_{m+1}} = 1 - \sum_{j=1}^{m+1} \pi_{i_j}$$

by (B.3) and (B.5).

Conclusion For any two items x and y, the condition $\text{Cov}(Y_x, Y_y) = -\pi_{x|ak}\pi_{y|ak}$ on the multiple responses results in the special case of the multinomial distribution.

Appendix C

Derivation of Higher Moments for Multiple Responses - Chapter 4

In Chapter 4, we need to compute Var$(\tilde{\omega}_{xy|k})$ expressed by 10 terms in equation (4.22) on page 128. Only three of these 10 terms can be easily computed and the purpose of this part of the appendix is to compute the remaining 7 terms. First we compute the terms $\mathbb{E}X_x\bar{X}_yX_{10}$ and $\mathbb{E}X_y\bar{X}_xX_{10}$ of (4.22).
We begin with

$$\begin{aligned}
\mathbb{E}X_xX_yX_A &= \mathbb{E}(X_A + X_C)(X_B + X_C)X_A \\
&= \mathbb{E}X_A^2X_B + \mathbb{E}X_A^2X_C + \mathbb{E}X_AX_BX_C + \mathbb{E}X_AX_C^2 \\
&= (N_2\pi_A^2\pi_B + N_1\pi_A\pi_B) + (N_2\pi_A^2X_C + N_1\pi_A\pi_C) \\
&+ N_2\pi_A\pi_B\pi_C + (N_2\pi_A\pi_C^2 + N_1\pi_A\pi_C) \\
&= N_2(\pi_A^2\pi_B + \pi_A^2X_C + \pi_A\pi_B\pi_C + \pi_A\pi_C^2) + N_1(\pi_A\pi_B + \pi_A\pi_C + \pi_A\pi_C) \\
&= N_2\pi_x\pi_y\pi_A + N_1(\pi_A\pi_B + 2\pi_A\pi_C)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}X_xX_yX_C &= \mathbb{E}(X_A + X_C)(X_B + X_C)X_C \\
&= \mathbb{E}X_AX_BX_C + \mathbb{E}X_AX_C^2 + \mathbb{E}X_BX_C^2 + \mathbb{E}X_C^3 \\
&= N_2\pi_A\pi_B\pi_C + (N_2\pi_A\pi_C^2 + N_1\pi_A\pi_C) \\
&+ (N_2\pi_B\pi_C^2 + N_1\pi_B\pi_C) + (N_2\pi_C^3 + 3N_1\pi_C^2 + N_0\pi_C) \\
&= N_2(\pi_A\pi_B\pi_C + \pi_A\pi_C^2 + \pi_BX_C^2 + \pi_C^3) + N_1(\pi_A\pi_C + \pi_B\pi_C + 3\pi_C^2) + N_0\pi_C \\
&= N_2\pi_x\pi_y\pi_C + N_1\pi_C(3\pi_C + \pi_A + \pi_B) + N_0\pi_C
\end{aligned}$$

Before we continue with $\mathbb{E}X_xX_yX_D$, we consider several types of symmetry. Consider two types of exchanging indices: (1) Exchanging items x and y ($x \to y$ and $y \to x$), (2-x) exchanging positive with negative responses for item x ($X_x \leftrightarrow \bar{X}_x$). These operations can also be regarded as transformations forming a transformation group. The pairwise observations change as follows under (1): $X_A \to X_B$, $X_C \to X_C$, $X_D \to X_D$, under (2-x): $X_A \to X_D$, $X_B \to X_C$, under (2-y): $X_B \to X_D$, $X_A \to X_C$, and under (2-x)$\circ$(2-y) : $X_A \to X_B$, $X_C \to X_D$, where $\circ$ denotes the operator executing two transformations.

We do not need to compute $\mathbb{E}X_xX_yX_B$ directly, but only note that $X_xX_yX_B$ can be obtained by applying (1) to $X_xX_yX_A$. We have $\mathbb{E}X_xX_yX_A = N_2\pi_x\pi_y\pi_A + N_1(\pi_A\pi_B + 2\pi_A\pi_C)$, hence, $\mathbb{E}X_xX_yX_B = N_2\pi_x\pi_y\pi_B + N_1(\pi_A\pi_B + 2\pi_B\pi_C)$. Now we can compute

$$\begin{aligned}
&\mathbb{E}X_xX_yX_D \\
&= \mathbb{E}(X_A + X_C)(X_B + X_C)(n - X_A - X_B - X_C) \\
&= n\mathbb{E}X_xX_y - \mathbb{E}X_xX_yX_A - \mathbb{E}X_xX_yX_B - \mathbb{E}X_xX_yX_C \\
&= (N_2 + 2N_1)\pi_x\pi_y + (N_1 + N_0)\pi_C - N_2\pi_x\pi_y\pi_A - N_1(\pi_A\pi_B + 2\pi_A\pi_C)
\end{aligned}$$

$$
\begin{aligned}
&- N_2\pi_x\pi_y\pi_B - N_1(\pi_A\pi_B + 2\pi_B\pi_C) - N_2\pi_x\pi_y\pi_C - N_1\pi_C(3\pi_C + \pi_A + \pi_B) + N_0\pi_C \\
&= N_2(\pi_x\pi_y - \pi_x\pi_y\pi_A - \pi_x\pi_y\pi_B - \pi_x\pi_y\pi_C) \\
&+ N_1\{2\pi_x\pi_y + \pi_C - \pi_A\pi_B - 2\pi_A\pi_C - \pi_A\pi_B - 2\pi_B\pi_C - \pi_C(3\pi_C + \pi_A + \pi_B)\} \\
&= N_2\pi_x\pi_y\pi_D + N_1(2\pi_x\pi_y - 2\pi_A\pi_B + \pi_C - 3\pi_A\pi_C - 3\pi_B\pi_C - 3\pi_C^2) \\
&= N_2\pi_x\pi_y\pi_D + N_1(2\pi_x\pi_y - 2\pi_x\pi_y + \pi_C - \pi_A\pi_C - \pi_B\pi_C - \pi_C^2) \\
&= N_2\pi_x\pi_y\pi_D + N_1\pi_C(1 - \pi_A - \pi_B - \pi_C) = N_2\pi_x\pi_y\pi_D + N_1\pi_C\pi_D.
\end{aligned}
$$

We summarise

$$
\begin{aligned}
\mathbb{E}X_xX_yX_A &= N_2\pi_x\pi_y\pi_A + N_1(\pi_A\pi_B + 2\pi_A\pi_C) \\
\mathbb{E}X_xX_yX_B &= N_2\pi_x\pi_y\pi_B + N_1(\pi_A\pi_B + 2\pi_B\pi_C) \\
\mathbb{E}X_xX_yX_C &= N_2\pi_x\pi_y\pi_C + N_1\pi_C(3\pi_C + \pi_A + \pi_B) + N_0\pi_C \\
\mathbb{E}X_xX_yX_D &= N_2\pi_x\pi_y\pi_D + N_1\pi_C\pi_D
\end{aligned} \tag{C.1}
$$

Now we obtain $\mathbb{E}X_x\bar{X}_yX_A$ from $\mathbb{E}X_xX_yX_C$ by (2-y)

$$
\begin{aligned}
\mathbb{E}X_x\bar{X}_yX_A &= N_2\pi_x\pi_y\pi_C + N_1\pi_A(3\pi_A + \pi_C + \pi_D) + N_0\pi_C \\
&= N_2\pi_x\pi_y\pi_C + N_1(3\pi_A^2 + \pi_A\pi_C + \pi_A(1 - \pi_A - \pi_B - \pi_C)) + N_0\pi_C \\
&= N_2\pi_x\pi_y\pi_C + N_1(2\pi_A^2 + \pi_A - \pi_A\pi_B) + N_0\pi_C
\end{aligned}
$$

and $\mathbb{E}X_x\bar{X}_yX_B$ from $\mathbb{E}X_xX_yX_D$ also by (2-y):

$$
\mathbb{E}X_x\bar{X}_yX_B = N_2\pi_x\bar{\pi}_y\pi_B + N_1\pi_A\pi_B.
$$

Using (1), we also easily obtain $\mathbb{E}\bar{X}_x X_y X_A$ and $\mathbb{E}\bar{X}_x X_y X_B$. Next, we compute

$$\begin{aligned}
&\mathbb{E}X_x^2 X_y = \mathbb{E}(X_A + X_C)^2(X_B + X_C) = \mathbb{E}(X_A^2 + X_C^2 + 2X_A X_C)(X_B + X_C) \\
&= \mathbb{E}X_A^2 X_B + \mathbb{E}X_A^2 X_C + \mathbb{E}X_B X_C^2 + \mathbb{E}X_C^3 + 2\mathbb{E}X_A X_B X_C + 2\mathbb{E}X_A X_C^2 \\
&= (N_2\pi_A^2\pi_B + N_1\pi_A\pi_B) + (N_2\pi_A^2\pi_C + N_1\pi_A\pi_C) + (N_2\pi_B\pi_C^2 + N_1\pi_B\pi_C) \\
&+ (N_2\pi_C^3 + 3N_1\pi_C^2 + N_0\pi_C) + 2N_2\pi_A\pi_B\pi_C + 2(N_2\pi_A\pi_C^2 + N_1\pi_A\pi_C) \\
&= N_2\left\{\pi_A^2\pi_B + \pi_A^2\pi_C + \pi_B\pi_C^2 + \pi_C^3 + \pi_A^2\pi_C + 2\pi_A\pi_B\pi_C + 2\pi_B\pi_C^2\right\} \\
&+ N_1\left\{\pi_A\pi_B + 3\pi_A\pi_C + \pi_B\pi_C + 3\pi_C^2\right\} + N_0\pi_C \\
&= N_2(\pi_A^2 + \pi_C^2 + 2\pi_A\pi_C)(\pi_B + \pi_C) + N_1(\pi_A + \pi_C)(\pi_B + 3\pi_C) + N_0\pi_C \\
&= N_2\pi_x^2\pi_y + N_1\pi_x(\pi_B + 3\pi_C) + N_0\pi_C
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}X_x^2 X_y^2 \\
&= \mathbb{E}(X_A + X_C)^2(X_B + X_C)^2 = \mathbb{E}(X_A^2 + X_C^2 + 2X_A X_C)(X_B^2 + X_C^2 + 2X_B X_C) \\
&= \mathbb{E}X_A^2 X_B^2 + \mathbb{E}X_A^2 X_C^2 + 2\mathbb{E}X_A^2 X_B X_C + \mathbb{E}X_B^2 X_C^2 + \mathbb{E}X_C^4 + 2\mathbb{E}X_B X_C^3 \\
&+ 2\mathbb{E}X_A X_B^2 X_C + 2\mathbb{E}X_A X_C^3 + 4\mathbb{E}X_A X_B X_C^2 \\
&= \{N_3\pi_A^2\pi_B^2 + N_2(\pi_A^2\pi_B + \pi_A\pi_B^2) + N_1\pi_A\pi_B\} \\
&+ \{N_3\pi_A^2\pi_C^2 + N_2(\pi_A^2\pi_C + \pi_A\pi_C^2) + N_1\pi_A\pi_C\} + 2\{N_3\pi_A^2\pi_B\pi_C + N_2\pi_A\pi_B\pi_C\} \\
&+ \{N_3\pi_B^2\pi_C^2 + N_2(\pi_B^2\pi_C + \pi_B\pi_C^2) + N_1\pi_B\pi_C\} + \{N_3\pi_C^4 + 6N_2\pi_C^3 + 7N_1\pi_C^2 + N_0\pi_C\} \\
&+ 2\{N_3\pi_B\pi_C^3 + 3N_2\pi_B\pi_C^2 + N_1\pi_B\pi_C\} + 2\{N_3\pi_A\pi_B^2\pi_C + N_2\pi_A\pi_B\pi_C\} \\
&+ 2\{N_3\pi_A\pi_C^3 + 3N_2\pi_A\pi_C^2 + N_1\pi_A\pi_C\} + 4\{N_3\pi_A\pi_B\pi_C^2 + N_2\pi_A\pi_B\pi_C\} \\
&= N_3\{\pi_A^2\pi_B^2 + \pi_A^2\pi_C^2 + 2\pi_A^2\pi_B\pi_C + \pi_B^2\pi_C^2 + \pi_C^4 + 2\pi_B\pi_C^3 \\
&+ 2\pi_A\pi_B^2\pi_C + 2\pi_A\pi_C^3 + 4\pi_A\pi_B\pi_C^2\} + N_2\{\pi_A^2\pi_B + \pi_A\pi_B^2 + \pi_A^2\pi_C + \pi_A\pi_C^2
\end{aligned}$$

$$
\begin{aligned}
&+ 2\pi_A\pi_B\pi_C + \pi_B^2\pi_C + \pi_B\pi_C^2 + 6\pi_C^3 + 6\pi_B\pi_C^2 + 2\pi_A\pi_B\pi_C + 6\pi_A\pi_C^2 + 4\pi_A\pi_B\pi_C\} \\
&+ N_1\{\pi_A\pi_B + \pi_A\pi_C + \pi_B\pi_C + 7\pi_C^2 + 2\pi_B\pi_C + 2\pi_A\pi_C\} + N_0\pi_C \\
&= N_3\{(\pi_A^2 + \pi_C^2 + 2\pi_A\pi_C)(\pi_B^2 + \pi_C^2 + 2\pi_B\pi_C\} \\
&+ N_1\{7\pi_C^2 + 3\pi_A\pi_C + 3\pi_B\pi_C + \pi_A\pi_B\} + N_0\pi_C \\
&+ N_2\{6\pi_C^3 + 7\pi_B\pi_C^2 + 7\pi_A\pi_C^2 + \pi_B^2\pi_C + 8\pi_A\pi_B\pi_C + \pi_A^2\pi_C + \pi_A^2\pi_B + \pi_A\pi_B^2\} \\
&= N_3\pi_x^2\pi_y^2 + N_1\{7\pi_C^2 + 3\pi_A\pi_C + 3\pi_B\pi_C + \pi_A\pi_B\} + N_0\pi_C \\
&+ N_2(\pi_A + \pi_B)(\pi_B + \pi_C)(\pi_A + \pi_B + 6\pi_C) \\
&= N_3\pi_x^2\pi_y^2 + N_2\pi_x\pi_y(\pi_A + \pi_B + 6\pi_C) + N_1\{7\pi_C^2 + 3\pi_A\pi_C + 3\pi_B\pi_C + \pi_A\pi_B\} + N_0\pi_C.
\end{aligned}
$$

The term $\mathbb{E}X_x^2\bar{X}_y^2$ is computed from $\mathbb{E}X_x^2X_y^2$ by (2-y)

$$
\begin{aligned}
&\mathbb{E}X_x^2\bar{X}_y^2 \\
&= N_3\pi_x\bar{\pi}_y^2 + N_2\pi_x\bar{\pi}_y(\pi_C + \pi_D + 6\pi_A) + N_1\{\pi_x\bar{\pi}_y + 6\pi_A^2 + 2\pi_A(\pi_C + \pi_D)\} + N_0\pi_A \\
&= N_3\pi_x\bar{\pi}_y^2 + N_2\pi_x\bar{\pi}_y(\pi_C + 1 - \pi_A - \pi_B - \pi_C + 6\pi_A) + N_0\pi_A \\
&+ N_1\{\pi_x\bar{\pi}_y + 6\pi_A^2 + 2\pi_A\pi_C + 2\pi_A - 2\pi_A^2 - 2\pi_A\pi_B - 2\pi_A\pi_C\} \\
&= N_3\pi_x\bar{\pi}_y^2 + N_2\pi_x\bar{\pi}_y(1 - \pi_B + 5\pi_A) + N_1(\pi_x\bar{\pi}_y + 4\pi_A^2 + 2\pi_A - 2\pi_A\pi_B) + N_0\pi_A.
\end{aligned}
$$

Summarising, we can write

$$
\begin{aligned}
\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y &= \mathbb{E}X_xX_y(n - X_x)(n - X_y) \\
&= n^2\mathbb{E}X_xX_y - n\mathbb{E}X_x^2X_y - n\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y^2. \qquad \text{(C.2)}
\end{aligned}
$$

For a better overview, we do not compute $\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y$ at once, but only for the terms with factors N_3, N_2, N_1, N_0 separately. Let$(\cdot)|_{N_i}$ denote the terms of $(\cdot)$ with factor N_i, for example $\mathbb{E}X_x^2X_y^2|_{N_3} = \pi_x^2\pi_y^2$.

Using (C.2) and (4.24) we collect the following terms for $\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y$

$$
\begin{aligned}
&\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_3} \\
&= (n^2\mathbb{E}X_xX_y - n\mathbb{E}X_x^2X_y - n\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y^2)|_{N_3} \\
&= \{(N_3 + 5N_2 + 4N_1)\mathbb{E}X_xX_y|_{N_1} + (N_2 + 3N_1 + N_0)\mathbb{E}X_xX_y|_{N_0} + \mathbb{E}X_x^2X_y^2 \\
&- 2((N_3 + 3N_2)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_2} + (N_2 + 2N_1)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_1} \\
&+ (N_1 + N_0)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_0})\}|_{N_3} \\
&= \mathbb{E}X_xX_y|_{N1} - \mathbb{E}X_x^2X_y|_{N_2} - \mathbb{E}X_xX_y^2|_{N_2} + \mathbb{E}X_x^2X_y^2|_{N_3} \\
&= \pi_x\pi_y - \pi_x^2\pi_y - \pi_x\pi_y^2 + \pi_x^2\pi_y^2 = \pi_x\pi_y(1 - \pi_x - \pi_y + \pi_x\pi_y) \\
&= \pi_x\pi_y\bar{\pi}_x\bar{\pi}_y
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_2} \\
&= (n^2\mathbb{E}X_xX_y - n\mathbb{E}X_x^2X_y - n\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y^2)|_{N_2} \\
&= \{(N_3 + 5N_2 + 4N_1)\mathbb{E}X_xX_y|_{N_1} + (N_2 + 3N_1 + N0)\mathbb{E}X_xX_y|_{N_0} + \mathbb{E}X_x^2X_y^2 \\
&- 2((N_3 + 3N_2)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_2} + (N_2 + 2N_1)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_1} \\
&+ (N_1 + N_0)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_0})\}|_{N_2} \\
&= 5\mathbb{E}X_xX_y|_{N_1} + \mathbb{E}X_xX_y|_{N_0} + \mathbb{E}X_x^2X_y^2|_{N_2} - 3\{\mathbb{E}X_x^2X_y|_{N_2} + \mathbb{E}X_xX_y^2|_{N_2}\} \\
&- \{\mathbb{E}X_x^2X_y|_{N_1} + \mathbb{E}X_xX_y^2|_{N_1}\} \\
&= 5\pi_x\pi_y + \pi_C + \pi_x\pi_y(\pi_A + \pi_B + 6\pi_C) - 3\{\pi_x^2\pi_y + \pi_x\pi_y^2\} \\
&- \{(3\pi_C^2 + 3\pi_A\pi_C + \pi_A\pi_B + \pi_B\pi_C) + (3\pi_C^2 + 3\pi_B\pi_C + \pi_A\pi_B + \pi_A\pi_C)\} \\
&= \pi_x\pi_y\{5 + \pi_A + \pi_B + 6\pi_C - 3\pi_x - 3\pi_y\} - 6\pi_C^2 - 2\pi_A\pi_B - 4\pi_A\pi_C - 4\pi_B\pi_C + \pi_C \\
&= \pi_x\pi_y\{5 - 2\pi_A - 2\pi_B\} - 6\pi_C^2 - 2\pi_A\pi_B - 4\pi_A\pi_C - 4\pi_B\pi_C + \pi_C
\end{aligned}
$$

$$= -2\pi_x\pi_y(\pi_A + \pi_B) + (\pi_x + \pi_y)(\pi_A + \pi_B) - (\pi_x + \pi_y)(\pi_A + \pi_B)$$

$$- 6\pi_C^2 - 2\pi_A\pi_B - 4\pi_A\pi_C - 4\pi_B\pi_C + 5\pi_x\pi_y + \pi_C$$

$$= (\pi_x + \pi_y - 2\pi_x\pi_y)(\pi_A + \pi_B) + 1/2(\pi_x + \pi_y - 2\pi_x\pi_y) - (\pi_x + \pi_y)(\pi_A + \pi_B)$$

$$- 6\pi_C^2 - 2\pi_A\pi_B - 4\pi_A\pi_C - 4\pi_B\pi_C + 5\pi_x\pi_y + \pi_C - 1/2(\pi_x + \pi_y - 2\pi_x\pi_y)$$

$$= \frac{1}{2}\{(\pi_x(1-\pi_y) + (1-\pi_x)\pi_y)(2\pi_A + 2\pi_B + 1) - 2(\pi_x + \pi_y)(\pi_A + \pi_B)$$

$$- 12\pi_C^2 - 4\pi_A\pi_B - 8\pi_A\pi_C - 8\pi_B\pi_C + 10\pi_x\pi_y + 2\pi_C - (\pi_x + \pi_y - 2\pi_x\pi_y)\}$$

$$= \frac{1}{2}\{(\pi_x\bar{\pi}_y + \bar{\pi}_x\pi_y)(2\pi_A + 2\pi_B + 1) - 2(\pi_A + \pi_B + 2\pi_C)(\pi_A + \pi_B)$$

$$- 12(\pi_C^2 + \pi_A\pi_B + \pi_A\pi_C + \pi_B\pi_C) + 8\pi_A\pi_B + 4\pi_A\pi_C + 4\pi_B\pi_C + 12\pi_x\pi_y + 2\pi_C - (\pi_x + \pi_y)\}$$

$$= \frac{1}{2}\{(\pi_x\bar{\pi}_y + \bar{\pi}_x\pi_y)(2\pi_A + 2\pi_B + 1)$$

$$- 2(\pi_A^2 + 2\pi_A\pi_B + \pi_B^2 + 2\pi_A\pi_C + 2\pi_B\pi_C)$$

$$- 12\pi_x\pi_y + 12\pi_x\pi_y + 8\pi_A\pi_B + 4\pi_A\pi_C + 4\pi_B\pi_C + 2\pi_C - (\pi_A + \pi_B + 2\pi_C)\}$$

$$= \frac{1}{2}\{(\pi_x\bar{\pi}_y + \bar{\pi}_x\pi_y)(2\pi_A + 2\pi_B + 1)$$

$$- 2\pi_A^2 - 2\pi_B^2 + 4\pi_A\pi_B - \pi_A - \pi_B + 4\pi_A\pi_C - 4\pi_A\pi_C + 4\pi_B\pi_C - 4\pi_B\pi_C\}$$

$$= \frac{1}{2}\{(\pi_x\bar{\pi}_y + \bar{\pi}_x\pi_y)(2\pi_A + 2\pi_B + 1) - 2(\pi_A - \pi_B)^2 - (\pi_A + \pi_B)\}$$

$$\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_1}$$

$$= (n^2\mathbb{E}X_xX_y - n\mathbb{E}X_x^2X_y - n\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y^2)|_{N_1}$$

$$= \{(N_3 + 5N_2 + 4N_1)\mathbb{E}X_xX_y|_{N_1} + (N_2 + 3N_1 + N_0)\mathbb{E}X_xX_y|_{N_0} + \mathbb{E}X_x^2X_y^2$$

$$- 2((N_3 + 3N_2)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_2} + (N_2 + 2N_1)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_1}$$

$$+ (N_1 + N_0)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_0})\}|_{N_1}$$

$$
\begin{aligned}
&= 4\mathbb{E}X_xX_y|_{N1} + 3\mathbb{E}X_xX_y|_{N0} + \mathbb{E}X_x^2X_y^2|_{N_1} - 2\{\mathbb{E}X_x^2X_y|_{N_2} + \mathbb{E}X_xX_y^2|_{N_2}\} \\
&- \{\mathbb{E}X_x^2X_y|_{N_0} + \mathbb{E}X_xX_y^2|_{N_0}\} \\
&= 4(\pi_A\pi_B + \pi_A\pi_C + \pi_B\pi_C + \pi_C^2) + 3\pi_C + (7\pi_C^2 + 3\pi_B\pi_C + 3\pi_A\pi_C + \pi_A\pi_B) \\
&- 2\{(3\pi_C^2 + \pi_B\pi_C + 3\pi_A\pi_C + \pi_A\pi_B) + (3\pi_C^2 + 3\pi_B\pi_C + \pi_A\pi_C + \pi_A\pi_B)\} \\
&- (\pi_C + \pi_C) \\
&= \pi_A\pi_B + \pi_C - \pi_A\pi_C - \pi_B\pi_C - \pi_C^2 \\
&= \pi_x\bar{\pi}_y - \pi_A = \pi_y\bar{\pi}_x - \pi_B
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}X_xX_y\bar{X}_x\bar{X}_y|_{N_0} &= (n^2\mathbb{E}X_xX_y - n\mathbb{E}X_x^2X_y - n\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y^2)|_{N_0} \\
&= \{(N_3 + 5N_2 + 4N_1)\mathbb{E}X_xX_y|_{N_1} + (N_2 + 3N_1 + N_0)\mathbb{E}X_xX_y|_{N_0} + \mathbb{E}X_x^2X_y^2 \\
&- 2((N_3 + 3N_2)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_2} + (N_2 + 2N_1)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_1} \\
&+ (N_1 + N_0)(\mathbb{E}X_xX_y^2 + \mathbb{E}X_x^2X_y)|_{N_0})\}|_{N_0} \\
&= \mathbb{E}X_xX_y|_{N_0} + \mathbb{E}X_x^2X_y^2|_{N_0} - \{\mathbb{E}X_x^2X_y|_{N_0} + \mathbb{E}X_xX_y^2|_{N_0}\} \\
&= \pi_C + \pi_C - (\pi_C + \pi_C) = 0.
\end{aligned}
$$

Appendix D

Computation of Asymptotic Variance for Model I of New MH estimator - Chapter 4

In Subsection 4.5.2 of Chapter 4 on page 135, we consider the asymptotic variance under the large-stratum limiting model of the newly proposed Mantel-Haenszel estimator $\tilde{\Psi}_{xy}$. In this part of the appendix, we use the delta method to compute this asymptotic variance.

We can write $\tilde{\omega}_{xy|k}/N$ as

$$\tilde{\omega}_{xy|k}/N = \frac{n_k}{N}\frac{n_k}{n'_k}\left\{\left(\frac{X_{x|k}}{n_k}\frac{\bar{X}_{y|k}}{n_k} + \frac{1}{n_k}\frac{X_{A|k}}{n_k}\right) - \Psi_{xy}\left(\frac{X_{y|k}}{n_k}\frac{\bar{X}_{x|k}}{n_k} + \frac{1}{n_k}\frac{X_{A|k}}{n_k}\right)\right\}$$

and define $g(\mathbf{p})$ as

$$\begin{aligned} g(\mathbf{p}_k) &= \alpha_k\left\{\frac{X_{x|k}}{n_k}\frac{\bar{X}_{y|k}}{n_k} - \Psi_{xy}\frac{X_{y|k}}{n_k}\frac{\bar{X}_{x|k}}{n_k}\right\} \\ &= \alpha_k\left\{(p_{A|k}+p_{C|k})(1-p_{B|k}-p_{C|k}) - \Psi_{xy}(p_{B|k}+p_{C|k})(1-p_{A|k}-p_{C|k})\right\}, \end{aligned}$$

where the sample proportions are defined as $\mathbf{p}_k = (p_A, p_B, p_C)$ with $p = \frac{X}{n}$. In the same way, we define $\boldsymbol{\pi}_k := (\pi_{A|k}, \pi_{B|k}, \pi_{C|k})$.

According to the multivariate C.L.T. (Theorem 2.8.6 on page 75), $\sqrt{N}(\mathbf{p}_k - \boldsymbol{\pi}_k) \rightarrow_d N(\mathbf{0}, \boldsymbol{\Sigma}_k)$ with

$$\boldsymbol{\Sigma}_k = (\Sigma)_{i,j=A}^{C} = \alpha_k^{-1}\{\text{Diag}(\boldsymbol{\pi}_k) - \boldsymbol{\pi}_k\boldsymbol{\pi}_k^T\}.$$

The random variables $\sqrt{N} \cdot \boldsymbol{p}_k$ and $\sqrt{N} \cdot (\tilde{\omega}_{xy|k}/N)$ consists of two summands, each a product of the sample proportions (one has additional factor Ψ), and factors α_k and $\frac{n_k}{N}\frac{n_k}{n'_k}$. Because $\frac{n_k}{N}\frac{n_k}{n'_k}$ converges to α_k, we conclude that the limiting normal distributions of the summands of both expressions are identical. It follows that $\sqrt{N} \cdot \boldsymbol{p}_k$ and $\sqrt{N} \cdot (\tilde{\omega}_{xy|k}/N)$ also have the same limiting normal distribution. In the same way as we wrote $g(\mathbf{p}_k)$, we can write

$$\begin{aligned} g(\boldsymbol{\pi}_k) &:= \alpha_k \left\{\pi_{x|k}\bar{\pi}_{y|k} - \Psi_{xy}\pi_{y|k}\bar{\pi}_{x|k}\right\} \\ &= \alpha_k \left\{(\pi_{A|k} + \pi_{C|k})(1 - \pi_{B|k} - \pi_{C|k}) - \Psi_{xy}(\pi_{B|k} + \pi_{C|k})(1 - \pi_{A|k} - \pi_{C|k})\right\}. \end{aligned}$$

Clearly, $g(\mathbf{p}) \rightarrow_p g(\boldsymbol{\pi}_k)[= \mathbb{E}\, g(\mathbf{p})]$, because $\mathbf{p} \rightarrow_p \boldsymbol{\pi}_k$, but also $\tilde{\omega}_{xy|k}/N \rightarrow_p g(\boldsymbol{\pi}_k)$.

We apply now the delta method (Theorem 2.8.4 on page 75) to $g(\mathbf{p}_k)$. The delta method says that $\sqrt{N}\{g(\mathbf{p}_k) - g(\boldsymbol{\pi}_k)\}$ is asymptotically normally distributed with mean zero and variance $V_g = \mathbf{B}^T\boldsymbol{\Sigma}_k\mathbf{B}$, where $\mathbf{B} = \frac{\partial g}{\partial \mathbf{p}}$ is the partial derivative matrix evaluated at $\boldsymbol{\pi}_k$.

Next we compute the derivatives. For convenience, we write $\partial g/\partial \pi$ for $\partial g/\partial p|_{p=\pi}$. We compute

$$\frac{\partial g}{\partial \pi_{A|k}} = \alpha_k \left[\bar{\pi}_{y|k} + \Psi\pi_{y|k}\right]$$

$$\frac{\partial g}{\partial \pi_{B|k}} = \qquad -\alpha_k \left[\pi_{x|k} + \Psi\bar{\pi}_{x|k}\right]$$
$$\frac{\partial g}{\partial \pi_{C|k}} = \alpha_k \left[\bar{\pi}_{y|k} + \Psi\pi_{y|k}\right] \qquad -\alpha_k \left[\pi_{x|k} + \Psi\bar{\pi}_{x|k}\right].$$

Now we write $V_g = \frac{\partial g}{\partial \pi}\mathbf{\Sigma}_k(\frac{\partial g}{\partial \pi})^T$ as

$$V_g = \sum_{i,j=A}^{C} \frac{\partial g}{\partial \pi_{i|k}}\frac{\partial g}{\partial \pi_{j|k}}\Sigma_{ij|k} = \sum_i (\pi_i - \pi_i^2)\left(\frac{\partial g}{\partial \pi_{i|k}}\right)^2 - \sum_{i=A}^{C}\sum_{j\neq i}\pi_i\pi_j\frac{\partial g}{\partial \pi_{i|k}}\frac{\partial g}{\partial \pi_{j|k}}$$
$$= \left\{\pi_{A|k}\left(\frac{\partial g}{\partial \pi_{A|k}}\right)^2 + \pi_{B|k}\left(\frac{\partial g}{\partial \pi_{B|k}}\right)^2 + \pi_{C|k}\left(\frac{\partial g}{\partial \pi_{C|k}}\right)^2\right\}$$
$$- \left\{\pi_{A|k}\frac{\partial g}{\partial \pi_{A|k}} + \pi_{B|k}\frac{\partial g}{\partial \pi_{B|k}} + \pi_{C|k}\frac{\partial g}{\partial \pi_{C|k}}\right\}^2$$
$$= \alpha_k\left\{\pi_A\left[\bar{\pi}_y + \Psi\pi_y\right]^2 + \pi_B\left[\pi_x + \Psi\bar{\pi}_x\right]^2 + \pi_C\left[(\bar{\pi}_y + \Psi\pi_y) - (\pi_x + \Psi\bar{\pi}_x)\right]^2\right\}$$
$$- \alpha_k\left\{\pi_A\left[\bar{\pi}_y + \Psi\pi_y\right] + \pi_B\left[\pi_x + \Psi\bar{\pi}_x\right] + \pi_C\left[(\bar{\pi}_y + \Psi\pi_y) - (\pi_x + \Psi\bar{\pi}_x)\right]\right\}^2$$
$$= \alpha_k\{\pi_x[\bar{\pi}_y^2 + \Psi^2\pi_y^2 + 2\Psi\bar{\pi}_y\pi_y] + \pi_y[\pi_x^2 + \Psi^2\bar{\pi}_x^2 + 2\Psi\pi_x\bar{\pi}_x]$$
$$- 2\pi_C[\pi_x\bar{\pi}_y + \Psi^2\bar{\pi}_x\pi_y + \Psi\pi_x\pi_y + \Psi\bar{\pi}_x\bar{\pi}_y] - [(\pi_x\bar{\pi}_y - \Psi\bar{\pi}_x\pi_y) + (\Psi - 1)\pi_x\pi_y]^2\}.$$

Under the common odds ratio assumption $\pi_x\bar{\pi}_y - \Psi\bar{\pi}_x\pi_y = 0$ and $V_g = \alpha_k\{T_1 + \Psi^2 T_2 + 2\Psi T_3\}$, where

$$T_1 = \pi_x\bar{\pi}_y^2 + \pi_y\pi_x^2 - 2\pi_C\pi_x\bar{\pi}_y - \pi_x^2\pi_y^2$$
$$T_2 = \pi_y\bar{\pi}_x^2 + \pi_x\pi_y^2 - 2\pi_C\pi_y\bar{\pi}_x - \pi_x^2\pi_y^2$$
$$T_3 = \pi_x\pi_y\bar{\pi}_y + \pi_x\pi_y\bar{\pi}_x - \pi_C\bar{\pi}_x\bar{\pi}_y - \pi_C\pi_x\pi_y + \pi_x^2\pi_y^2.$$

We re-express T_1 as

$$T_1 = \pi_x\bar{\pi}_y^2 + \pi_y\pi_x^2 - 2\pi_C\pi_x\bar{\pi}_y - \pi_x^2\pi_y^2$$
$$= \pi_x\bar{\pi}_y^2 + \pi_x^2(-\bar{\pi}_y + 1) - 2\pi_C\pi_x\bar{\pi}_y - \pi_x^2\pi_y(-\bar{\pi}_y + 1)$$

$$= \pi_x\bar{\pi}_y(\bar{\pi}_y - \pi_x - 2\pi_C + \pi_x\pi_y) + \pi_x^2(1-\pi_y)$$
$$= \pi_x\bar{\pi}_y(\bar{\pi}_y - 2\pi_C + \pi_x\pi_y),$$

similarly $T_2 = \pi_y\bar{\pi}_x(\bar{\pi}_x - 2\pi_C + \pi_x\pi_y)$. Let us write T_3 as

$$= \pi_x\pi_y\bar{\pi}_y + \pi_x\pi_y\bar{\pi}_x - \pi_C\bar{\pi}_x\bar{\pi}_y - \pi_C\pi_x\pi_y + \pi_x^2\pi_y^2$$
$$= (-\bar{\pi}_x+1)\pi_y\bar{\pi}_y + \pi_x\pi_y\bar{\pi}_x - 2\pi_C(-\bar{\pi}_x+1)\pi_y - \pi_C + \pi_C\pi_x + \pi_c\pi_y + (-\bar{\pi}_x+1)\pi_x\pi_y^2$$
$$= \bar{\pi}_x\pi_y(-\bar{\pi}_y + 2\pi_C - \pi_x\pi_y) + \pi_y\bar{\pi}_y - 2\pi_C\pi_y + \pi_C\pi_x + \pi_C\pi_y + \pi_x\pi_y^2 - \pi_C + \pi_x\bar{\pi}_x\pi_y$$
$$= \bar{\pi}_x\pi_y(-\bar{\pi}_y + 2\pi_C - \pi_x\pi_y) + \pi_y\bar{\pi}_y + \pi_C(\pi_x - \pi_y) + \pi_x\pi_y(\pi_y + \bar{\pi}_x) - \pi_C.$$

In the same way, we can express T_3 also as

$$\bar{\pi}_y\pi_x(-\bar{\pi}_x + 2\pi_C - \pi_x\pi_y) + \pi_x\bar{\pi}_x + \pi_C(\pi_y - \pi_x) + \pi_x\pi_y(\pi_x + \bar{\pi}_y) - \pi_C,$$

yielding

$$2\cdot T_3 = \bar{\pi}_x\pi_y(-\bar{\pi}_y+2\pi_C-\pi_x\pi_y)+\bar{\pi}_y\pi_x(-\bar{\pi}_x+2\pi_C-\pi_x\pi_y)+\pi_x\bar{\pi}_x+\pi_y\bar{\pi}_y+2(\pi_x\pi_y-\pi_C).$$

Under the common odds ratio assumption, we summarise

$$T_1 + \Psi^2 T_2 + 2\Psi T_3 = \pi_x\bar{\pi}_x + \pi_y\bar{\pi}_y + 2(\pi_x\pi_y - \pi_C).$$

Finally, we can write

$$\left[\lim_{N\to\infty} N\cdot \mathsf{Var}^a(\tilde{\omega}_{xy|k}/N) = \lim_{N\to\infty}\frac{1}{N}\mathsf{Var}^a(\tilde{\omega}_{xy|k}) =\right] V_g = \alpha_k\{\pi_x\bar{\pi}_x+\pi_y\bar{\pi}_y+2(\pi_x\pi_y-\pi_C)\}. \tag{D.1}$$

Appendix E

Normative Aging Study - Chapter 7

Table E.1: Normative Aging Study (NAS) - data set for all 682 men

subject	FBG	age	smk	wbc	crt	subject	FBG	age	smk	wbc	crt	subject	FBG	age	smk	wbc	crt
1	1	66	4	6.0	0.1360	2	1	75	4	4.5	0.2210	3	1	70	4	7.6	0.0681
4	3	67	4	8.6	0.4350	5	1	69	3	9.5	0.0245	6	1	67	1	6.1	0.0360
7	2	69	1	6.5	2.6730	8	1	74	4	4.4	0.1717	9	2	68	1	7.0	0.0810
10	1	66	1	6.5	0.1870	11	1	75	1	3.6	0.0520	12	1	68	4	7.6	0.1090
13	1	71	1	4.8	0.0980	14	1	67	4	7.4	0.2160	15	1	94	1	5.6	0.0100
16	2	76	4	6.1	0.2270	17	1	78	4	6.1	0.1420	18	1	74	1	5.1	0.1120
19	2	77	1	7.4	0.4100	20	3	71	4	4.8	0.1010	21	1	77	4	9.3	0.4570
22	2	86	1	7.5	0.5610	23	1	82	4	5.4	0.5400	24	1	67	4	4.4	0.0470
25	3	77	1	4.6	0.0108	26	2	82	1	5.5	0.1040	27	1	78	4	8.5	0.4160
28	1	76	4	6.0	0.4330	29	2	72	1	7.7	2.3130	30	2	81	4	5.6	0.0510
31	1	82	3	5.6	0.0250	32	1	74	1	8.9	0.1290	33	2	77	4	5.8	0.1650
34	3	62	4	7.9	1.3870	35	1	77	4	5.0	0.2300	36	2	63	4	7.0	0.4940
37	1	72	4	6.8	0.1450	38	2	82	4	8.4	0.0183	39	1	84	4	4.9	0.0400
40	1	78	4	5.5	0.0340	41	1	86	4	5.7	0.5440	42	2	77	1	6.9	0.2730
43	2	82	4	5.4	0.1110	44	1	76	4	5.3	0.1450	45	3	85	1	9.8	0.1987
46	1	70	1	5.4	0.2680	47	1	83	4	5.3	0.0990	48	2	77	4	6.6	0.0140
49	2	75	4	9.1	0.0059	50	2	69	4	6.1	0.1420	51	3	69	4	5.9	0.3580
52	2	71	1	6.1	0.0660	53	1	70	4	4.2	0.2400	54	1	69	1	7.5	0.0820
55	1	84	1	5.2	0.3390	56	1	70	4	5.6	0.1680	57	1	81	1	4.3	0.1090
58	1	70	4	4.9	0.9700	59	2	68	4	5.9	0.0190	60	1	80	4	5.7	0.1620
61	2	68	4	4.9	0.0860	62	1	74	1	6.2	0.2410	63	2	76	4	5.3	0.4060
64	1	89	4	5.7	0.0710	65	2	77	4	4.9	0.1170	66	3	69	4	5.1	1.8880
67	1	74	1	7.6	0.6500	68	3	74	4	12.4	0.0860	69	1	73	4	5.7	0.4970
70	3	69	1	6.9	0.0107	71	1	79	4	36.6	0.1330	72	1	78	1	4.8	0.0026
73	1	75	4	4.2	0.6700	74	2	77	4	6.3	0.1040	75	2	79	1	8.7	0.0360

subject	FBG	age	smk	wbc	crt	subject	FBG	age	smk	wbc	crt	subject	FBG	age	smk	wbc	crt
76	1	82	4	8.3	0.5820	77	1	78	1	4.2	0.2750	78	1	66	4	5.8	0.1280
79	1	68	1	5.8	0.3330	80	1	81	1	5.1	0.2460	81	2	75	3	8.3	0.5880
82	1	87	4	5.0	0.6550	83	1	78	4	6.0	0.3210	84	2	69	4	6.5	0.1690
85	1	73	1	6.1	0.0650	86	1	68	4	6.8	0.0610	87	1	79	4	7.9	0.1940
88	1	75	4	6.4	0.0570	89	2	81	1	6.5	0.1670	90	2	69	4	4.4	0.1570
91	1	71	4	4.8	0.1950	92	1	65	3	12.1	0.0153	93	1	74	1	6.3	0.1050
94	1	68	4	6.9	0.3360	95	1	73	4	7.8	0.0043	96	1	79	1	6.2	0.2830
97	1	77	4	7.2	0.0176	98	1	71	1	8.8	1.3970	99	2	74	1	7.8	0.1480
100	3	73	4	5.6	0.1120	101	3	68	4	6.5	0.1630	102	1	70	4	6.4	0.0610
103	2	69	4	5.6	0.1400	104	2	66	1	9.3	0.0970	105	3	76	1	6.6	0.0113
106	1	68	4	8.5	0.0073	107	1	66	4	5.4	0.0320	108	1	71	1	6.5	0.2290
109	1	71	4	7.7	0.0840	110	1	74	4	7.4	0.1010	111	1	66	4	7.3	0.0660
112	1	67	1	7.1	0.6340	113	1	83	4	8.6	5.5350	114	1	77	4	5.5	0.3720
115	1	78	1	5.7	0.1990	116	1	77	4	8.5	2.0750	117	3	72	4	4.3	0.2420
118	1	69	4	4.1	0.1580	119	1	72	4	5.2	0.0820	120	3	85	1	7.4	0.0810
121	3	74	4	8.8	0.2840	122	1	83	4	5.4	0.0490	123	2	79	4	7.7	0.3780
124	1	75	4	5.5	0.1150	125	2	64	1	8.4	0.6000	126	1	77	4	7.1	0.1710
127	1	80	1	5.4	0.0610	128	1	83	1	7.8	0.0389	129	1	73	4	4.2	0.1900
130	1	80	4	4.5	0.2240	131	1	80	1	5.1	0.2670	132	1	80	1	7.3	0.0580
133	2	83	1	6.4	1.4690	134	1	76	1	9.1	1.6740	135	3	86	1	4.2	0.6000
136	1	76	4	5.2	0.0680	137	3	71	4	4.6	0.0560	138	2	81	4	4.9	0.0980
139	3	76	4	8.8	0.1820	140	3	76	4	5.6	0.0147	141	3	79	4	4.1	0.2730
142	1	74	1	4.1	0.2890	143	1	86	4	15.1	0.8600	144	1	75	4	5.3	0.3010
145	3	78	1	6.5	0.3180	146	1	84	1	5.4	0.0890	147	3	69	4	10.8	0.4380
148	3	77	3	5.9	0.0211	149	1	81	4	5.4	0.6800	150	1	70	1	5.5	0.4770
151	2	74	1	8.1	1.4870	152	2	74	1	4.4	0.4200	153	1	68	4	4.8	0.1970
154	1	80	3	10.8	0.1410	155	1	83	4	8.3	0.2020	156	1	68	4	5.4	0.5010
157	2	63	4	5.5	0.2090	158	1	79	4	4.8	0.2040	159	3	78	4	6.5	0.1080
160	1	80	4	5.9	0.0418	161	1	74	4	6.6	0.1160	162	1	72	4	6.1	0.3830
163	2	84	4	7.0	0.2790	164	1	75	4	5.4	0.1910	165	1	67	3	4.8	0.0610
166	1	71	4	5.9	0.0058	167	1	69	4	5.3	0.2710	168	1	65	4	6.0	0.1440
169	2	73	4	8.0	0.3090	170	1	68	4	7.5	0.3030	171	1	89	4	10.4	0.1220
172	1	75	4	11.0	0.8480	173	3	62	4	3.3	0.0760	174	1	75	4	6.1	0.2500
175	1	81	1	5.9	0.1270	176	1	71	1	4.7	0.9500	177	1	82	4	11.1	17.2800
178	1	77	4	5.2	0.0760	179	1	75	3	3.4	0.4130	180	1	76	4	6.1	0.2500
181	1	75	1	5.3	0.0420	182	3	87	1	6.1	0.9490	183	1	69	4	9.3	0.3300
184	1	78	3	6.1	0.2520	185	1	86	1	6.2	0.4540	186	3	73	1	6.6	0.8750
187	1	71	4	4.8	0.2400	188	3	74	3	5.4	0.1430	189	2	82	4	6.0	0.2170
190	1	71	4	6.6	0.1480	191	1	68	4	4.4	0.4700	192	1	84	1	5.8	0.3530
193	3	74	4	7.2	0.2220	194	1	81	4	6.6	0.1660	195	3	75	1	4.6	0.1500
196	1	100	1	8.1	0.2390	197	1	80	4	7.5	0.6940	198	1	91	1	6.7	0.0551
199	2	68	1	5.8	0.1730	200	1	80	4	5.9	0.1080	201	3	70	4	6.4	0.5100
202	1	71	1	4.2	0.2270	203	1	67	4	5.2	0.1380	204	1	71	4	3.9	0.6840
205	3	68	4	7.2	0.3750	206	1	86	1	7.0	0.0590	207	1	79	4	6.7	0.2220
208	1	86	1	4.9	0.0120	209	1	76	4	6.8	0.3930	210	2	64	3	5.6	0.0550
211	1	89	4	7.0	0.0383	212	1	88	1	8.8	0.0850	213	1	89	1	10.1	0.2310
214	1	79	4	4.8	0.0980	215	1	81	4	8.2	2.4900	216	1	80	1	3.4	0.0101
217	1	81	4	5.3	0.0250	218	2	73	4	6.9	0.0660	219	1	65	4	4.7	0.1530
220	1	77	1	6.7	0.2100	221	1	65	1	5.3	0.0730	222	1	74	3	10.1	0.2330
223	1	84	1	5.4	0.6180	224	2	71	3	13.4	1.7440	225	1	87	4	6.1	0.0690
226	1	79	4	5.7	0.1760	227	3	78	1	7.3	0.1200	228	2	80	1	8.3	0.0223

subject	FBG	age	smk	wbc	crt	subject	FBG	age	smk	wbc	crt	subject	FBG	age	smk	wbc	crt
229	1	79	1	6.1	0.0880	230	1	72	4	4.5	0.6100	231	1	75	3	7.6	0.3840
232	1	82	1	7.2	0.2100	233	2	70	4	9.1	1.9260	234	1	76	4	8.0	0.0840
235	1	74	1	5.3	0.5170	236	1	75	4	7.3	0.5200	237	1	70	1	5.3	0.0790
238	1	86	4	8.8	0.0460	239	2	71	1	6.8	0.0530	240	2	86	1	8.7	0.9110
241	3	77	1	5.7	0.6800	242	1	77	4	5.8	0.1560	243	1	79	4	4.5	0.0850
244	3	84	1	5.0	0.0900	245	1	83	4	5.6	0.3100	246	1	83	1	4.1	0.1360
247	1	76	1	6.8	0.1160	248	2	71	4	14.6	8.6280	249	1	79	4	5.4	0.0660
250	1	66	1	4.9	0.1060	251	1	74	1	7.3	0.1370	252	1	78	4	8.9	0.0390
253	1	80	4	6.2	0.2950	254	2	68	4	5.4	0.0144	255	1	82	4	6.0	0.1660
256	3	68	4	5.6	0.3920	257	1	79	4	2.1	0.0600	258	3	69	4	9.9	0.3430
259	1	77	1	6.6	0.8710	260	2	82	4	8.8	0.6300	261	2	76	4	4.6	0.4600
262	1	77	1	4.2	0.8420	263	1	69	4	4.3	0.0150	264	1	61	3	12.0	0.3330
265	1	65	4	4.7	0.4250	266	2	76	4	5.7	0.0390	267	1	78	4	3.9	0.1470
268	2	70	4	5.6	0.0160	269	1	79	4	8.2	0.2220	270	1	72	4	6.4	0.0480
271	1	69	4	6.3	0.2470	272	1	70	4	6.4	0.2990	273	1	71	4	5.9	0.1320
274	1	69	4	4.7	0.2910	275	1	64	4	6.7	0.5890	276	1	79	4	3.6	0.6000
277	1	71	4	2.7	0.1140	278	2	72	1	6.4	0.4600	279	2	81	4	9.2	0.2150
280	2	75	3	10.1	0.5760	281	3	79	4	6.9	0.1260	282	2	73	3	6.8	0.1610
283	1	90	4	8.7	0.0022	284	1	89	4	7.7	0.1350	285	1	89	1	6.9	0.0990
286	3	70	4	6.2	1.8610	287	3	82	4	4.7	0.1030	288	3	70	4	4.7	0.0420
289	1	71	4	6.5	0.0440	290	1	79	4	5.5	0.8000	291	1	71	4	41.1	0.0640
292	1	73	1	5.3	0.0570	293	1	72	1	3.7	0.0760	294	3	75	4	4.3	0.0080
295	3	70	1	6.6	0.5400	296	3	79	4	6.0	0.0920	297	1	78	1	7.9	0.2160
298	1	76	4	7.6	0.2050	299	1	74	4	6.5	0.4080	300	1	66	4	7.0	0.2500
301	1	78	4	6.7	0.4710	302	3	78	4	5.3	0.3580	303	1	74	4	6.7	0.6710
304	2	77	4	4.9	0.0022	305	1	83	4	4.3	0.1967	306	1	74	1	4.9	0.0259
307	1	73	4	8.0	0.1760	308	2	73	4	6.8	0.0310	309	2	77	4	5.6	0.2260
310	1	78	4	7.1	1.2690	311	1	72	1	5.9	0.1340	312	2	75	4	4.8	0.1110
313	1	70	4	4.6	0.1650	314	1	76	1	9.0	0.4050	315	1	68	4	4.4	0.7170
316	1	75	4	4.7	0.0510	317	1	70	4	4.9	0.3400	318	1	69	4	7.1	0.1130
319	1	70	4	7.4	0.0720	320	3	80	4	7.8	0.0770	321	3	66	4	7.0	0.6480
322	3	80	4	4.8	0.0070	323	1	76	4	7.1	0.1970	324	1	69	1	6.3	0.1670
325	1	70	4	6.9	1.4700	326	1	77	4	10.7	0.3010	327	1	73	4	5.3	0.3160
328	1	63	4	4.1	0.0710	329	2	92	4	5.9	0.2214	330	1	68	4	3.5	0.0800
331	2	85	4	7.0	0.0185	332	1	75	4	8.9	3.7530	333	1	69	4	2.6	0.3110
334	3	71	4	4.6	0.0081	335	1	90	1	5.8	0.0990	336	2	73	4	5.6	0.0141
337	1	59	1	5.1	0.1330	338	3	71	4	8.4	0.0208	339	1	67	4	7.1	2.2090
340	1	66	1	5.3	0.3470	341	3	68	4	6.8	0.1610	342	1	76	1	8.6	0.4000
343	3	70	1	7.4	0.0550	344	1	69	1	5.7	0.0200	345	1	73	4	4.4	0.0098
346	1	78	1	4.2	0.3870	347	1	69	4	4.7	0.1050	348	3	72	4	5.7	0.3210
349	2	72	1	7.1	0.2070	350	1	69	4	5.7	0.0930	351	1	72	3	7.4	0.0144
352	3	74	4	4.6	0.0300	353	2	82	4	6.0	0.1710	354	1	73	1	6.2	0.0407
355	1	82	1	6.6	0.1060	356	1	70	4	4.9	0.1630	357	1	75	1	6.0	0.0810
358	2	78	4	3.7	0.4800	359	1	67	4	8.0	0.1630	360	1	73	4	5.7	0.1170
361	2	68	1	6.5	0.2700	362	1	78	4	6.4	0.3980	363	1	72	4	4.6	0.0790
364	1	78	1	7.4	0.0290	365	1	64	4	5.7	0.1380	366	1	75	4	9.2	0.2700
367	1	68	4	5.4	0.0690	368	1	78	4	7.7	0.5210	369	1	70	1	4.4	0.0620
370	1	77	4	7.0	0.4030	371	1	71	4	7.7	1.8430	372	1	75	1	6.9	0.3640
373	1	73	4	10.0	1.5940	374	1	68	4	16.4	0.0370	375	1	65	1	5.1	0.0480
376	1	65	4	6.0	0.1260	377	2	75	4	6.9	0.2450	378	3	74	4	5.9	0.2290
379	1	70	1	20.6	0.0220	380	1	76	1	6.5	0.1710	381	1	69	1	7.4	0.1520

subject	FBG	age	smk	wbc	crt	subject	FBG	age	smk	wbc	crt	subject	FBG	age	smk	wbc	crt
382	1	78	1	5.4	0.0053	383	1	62	4	5.5	0.2650	384	1	70	4	5.0	0.0950
385	1	77	1	6.3	0.0780	386	1	73	4	5.3	0.0830	387	1	71	1	5.6	0.0020
388	1	77	1	6.2	0.0320	389	1	89	4	6.1	0.2420	390	2	75	4	4.5	0.2250
391	1	80	4	5.3	0.5950	392	1	62	4	5.1	0.0950	393	3	66	4	7.0	6.7300
394	1	61	4	4.1	0.1200	395	1	80	4	4.7	0.0840	396	3	76	1	6.8	0.2880
397	1	73	4	6.4	0.2450	398	1	63	3	9.2	0.3470	399	1	71	4	5.3	0.3400
400	1	73	4	4.7	0.7400	401	1	75	1	5.0	0.0560	402	2	85	1	6.2	0.2360
403	1	80	4	6.3	0.3460	404	1	65	3	7.3	0.4270	405	1	70	4	8.7	0.0560
406	1	89	1	5.1	0.1310	407	1	70	1	6.7	0.2410	408	1	68	4	5.9	0.0330
409	3	69	1	5.2	0.3990	410	1	80	4	5.4	0.3420	411	1	79	4	6.7	0.1030
412	2	70	4	4.9	0.6300	413	1	67	4	4.2	0.0770	414	1	71	1	6.8	0.1380
415	1	69	4	9.1	0.7570	416	3	76	4	7.8	0.1720	417	1	79	4	3.8	0.0100
418	1	75	1	6.5	0.0397	419	1	70	4	6.1	0.9250	420	2	74	4	8.2	0.5690
421	1	78	4	6.1	0.3010	422	1	73	4	6.2	0.5770	423	2	70	4	5.8	0.0840
424	1	67	4	7.6	0.1700	425	1	91	1	5.3	0.1440	426	1	79	4	5.7	0.1920
427	1	72	4	5.5	0.1250	428	1	84	1	8.3	0.1500	429	1	76	4	5.6	0.1370
430	1	71	3	7.5	0.1130	431	3	71	1	8.5	2.2100	432	1	75	1	6.3	2.5210
433	2	79	4	9.4	0.2620	434	3	61	4	6.1	0.0830	435	1	82	1	4.4	0.2240
436	2	76	4	7.1	0.4030	437	1	68	4	8.3	0.3530	438	1	60	4	7.6	0.3280
439	1	86	4	4.8	0.2000	440	1	64	4	3.0	0.1000	441	1	74	4	7.7	0.7200
442	2	64	4	8.0	0.1020	443	1	73	4	6.6	0.1700	444	1	70	1	5.3	0.0750
445	2	75	1	7.2	0.1480	446	1	81	3	6.8	0.6520	447	1	66	4	10.0	0.1620
448	1	67	4	5.8	0.0770	449	2	78	4	4.0	0.0640	450	1	76	4	8.0	0.4720
451	1	65	1	4.1	0.0320	452	1	75	4	6.3	0.1230	453	2	70	4	7.2	0.5290
454	1	81	4	6.6	0.1060	455	1	76	4	7.0	0.8300	456	1	80	1	7.0	0.1260
457	1	68	4	4.5	0.7070	458	3	87	1	59.1	0.0180	459	1	75	1	2.2	0.3620
460	1	65	3	7.4	0.1840	461	3	70	4	6.2	0.2900	462	1	76	4	5.2	0.1510
463	1	61	4	4.9	0.1130	464	1	74	4	10.4	0.0297	465	1	63	3	6.1	0.1080
466	1	62	1	5.3	0.3030	467	1	74	4	7.4	0.2370	468	1	86	4	23.0	0.0400
469	1	74	4	4.3	0.1200	470	1	83	4	6.6	0.1260	471	1	67	1	5.6	0.3400
472	1	70	1	5.6	0.0025	473	1	69	4	5.1	0.7130	474	1	80	4	6.3	0.6570
475	1	80	1	5.8	0.1160	476	1	66	4	4.9	0.0750	477	1	66	1	6.7	0.0700
478	1	75	4	4.5	0.0660	479	1	80	4	10.2	0.6860	480	1	80	4	8.0	0.3430
481	1	71	1	5.2	0.5560	482	2	62	1	5.8	0.5370	483	3	76	1	4.6	0.3550
484	1	75	4	8.0	0.4360	485	1	76	4	7.5	0.1440	486	1	68	4	5.8	0.1500
487	2	78	4	7.6	0.1220	488	1	68	4	4.9	0.4300	489	2	68	3	11.0	0.6560
490	1	75	4	5.1	0.7700	491	1	68	4	5.9	0.0770	492	1	69	1	6.6	0.3490
493	1	69	4	4.6	0.4380	494	1	62	4	7.1	0.1770	495	1	78	1	11.2	0.7500
496	1	70	4	4.8	0.2290	497	1	70	4	6.0	0.6020	498	1	67	4	8.9	0.3190
499	1	76	4	4.8	0.3830	500	3	73	4	8.2	0.5480	501	1	80	4	6.1	0.2990
502	3	60	4	7.2	0.2630	503	1	86	1	5.9	0.0050	504	1	81	4	5.6	0.1500
505	1	72	1	4.8	0.0790	506	1	71	4	5.5	0.0630	507	3	66	4	6.6	0.6800
508	1	65	1	6.3	0.0088	509	1	80	1	5.8	0.0230	510	1	70	4	7.0	0.0990
511	1	77	1	5.9	0.5900	512	3	80	4	6.7	0.0071	513	1	64	3	8.5	0.5290
514	2	66	1	7.4	0.1910	515	1	71	4	8.1	0.7460	516	3	75	4	5.7	0.8000
517	1	76	1	4.6	0.6000	518	1	78	1	5.5	0.8350	519	3	74	1	6.4	0.3640
520	1	75	4	3.4	0.0710	521	3	80	4	3.6	0.2010	522	1	64	3	11.7	0.1620
523	1	62	4	6.7	0.9700	524	1	66	1	5.1	0.1060	525	2	67	4	6.3	0.3430
526	1	76	1	6.0	0.2160	527	3	62	3	10.4	0.1877	528	1	74	4	4.0	0.0106
529	1	78	1	3.1	0.0220	530	1	65	4	3.7	0.7030	531	2	79	4	8.6	0.0062
532	1	86	1	6.6	0.5260	533	1	84	1	5.9	0.3910	534	1	78	1	4.5	0.2630

subject	FBG	age	smk	wbc	crt	subject	FBG	age	smk	wbc	crt	subject	FBG	age	smk	wbc	crt
535	1	74	1	5.5	0.0850	536	2	76	4	4.8	0.2100	537	1	80	1	9.6	0.1670
538	2	77	1	5.3	0.1200	539	2	62	4	6.7	0.0310	540	3	66	4	7.1	0.4390
541	1	72	4	6.4	0.2070	542	1	67	4	4.5	0.1420	543	1	73	4	6.6	0.6000
544	1	82	4	6.4	0.3810	545	1	72	1	6.4	0.4880	546	3	76	4	9.2	0.8120
547	1	70	4	5.5	0.2190	548	1	89	4	8.2	0.0490	549	1	71	4	6.9	0.0750
550	1	67	1	3.0	0.4400	551	1	72	4	5.8	0.3130	552	2	72	4	7.8	0.6350
553	1	81	4	6.4	0.0038	554	1	65	1	6.1	0.1880	555	2	72	4	6.2	0.0147
556	1	68	4	5.4	0.1540	557	1	61	4	5.2	0.4200	558	1	65	1	7.1	0.2520
559	1	63	1	5.9	0.0560	560	2	63	4	7.7	0.1530	561	1	75	4	6.0	0.0610
562	2	68	4	4.5	0.1360	563	2	80	4	5.3	0.1360	564	2	76	3	1.8	0.1960
565	1	67	1	6.8	0.2580	566	1	74	1	5.7	0.6900	567	1	84	1	6.0	0.0032
568	1	69	4	7.7	1.7540	569	2	69	4	5.6	0.3060	570	1	59	1	6.9	0.3980
571	3	80	4	6.9	0.3290	572	1	80	1	4.6	0.0210	573	2	85	4	6.1	0.1030
574	2	70	1	4.7	0.2770	575	1	91	4	5.6	2.4520	576	1	69	4	6.5	0.4640
577	1	78	1	5.6	0.0730	578	1	72	4	3.8	0.0170	579	1	64	1	3.3	0.1830
580	1	79	4	8.2	0.2510	581	1	70	4	6.7	0.4600	582	1	73	3	5.8	0.0400
583	1	77	4	6.9	0.2290	584	1	70	4	5.7	0.1890	585	1	78	4	4.7	0.1780
586	1	75	1	6.0	0.1210	587	1	79	1	6.4	0.1720	588	1	71	4	7.1	0.1920
589	2	71	4	7.1	0.1820	590	3	76	4	8.4	0.0026	591	1	65	1	6.7	0.2480
592	3	67	4	8.2	0.5250	593	3	76	4	8.3	0.1780	594	3	73	4	6.7	1.4830
595	1	76	1	6.0	0.0038	596	1	64	4	6.7	0.3120	597	1	77	4	5.7	0.3520
598	1	75	1	5.9	0.3800	599	1	75	4	10.1	0.2120	600	2	80	4	6.8	4.0940
601	1	75	4	7.4	0.0450	602	1	77	4	5.9	0.4170	603	1	83	4	5.4	0.2620
604	1	67	4	4.7	0.0320	605	2	64	1	4.6	0.6510	606	1	82	4	5.1	0.0019
607	1	77	1	6.6	0.0950	608	1	64	3	4.7	0.0530	609	2	77	4	5.5	2.4330
610	1	67	4	4.7	0.0470	611	1	69	1	5.9	0.0680	612	1	73	1	5.4	0.0021
613	1	79	1	5.9	0.1470	614	1	82	1	3.8	0.2470	615	1	75	4	5.7	0.0540
616	1	72	1	11.3	0.1920	617	1	67	4	5.2	0.0360	618	1	80	4	4.6	0.4140
619	1	78	4	5.2	0.0135	620	1	74	1	6.1	0.1370	621	1	66	4	3.9	0.1140
622	3	61	4	6.6	0.0600	623	3	78	4	5.9	0.2240	624	3	76	1	8.6	0.2840
625	1	61	4	7.6	0.0200	626	1	72	4	6.2	0.0173	627	3	73	4	4.6	0.4700
628	1	64	4	5.1	0.1470	629	3	74	1	6.0	0.5080	630	2	73	4	6.3	0.4350
631	1	64	4	5.1	0.3390	632	3	74	1	5.3	0.2960	633	1	74	1	6.1	0.0880
634	1	70	4	6.8	0.1280	635	3	89	4	10.5	0.1220	636	1	73	4	7.6	0.4600
637	2	71	4	6.5	0.2590	638	1	64	3	10.1	0.2390	639	1	67	4	4.6	0.0470
640	3	73	1	5.1	0.1730	641	1	79	1	6.6	0.6280	642	3	57	4	6.4	0.2910
643	3	76	4	4.0	0.1810	644	1	66	4	9.8	2.8820	645	2	78	4	4.6	0.5900
646	1	73	4	8.5	0.1300	647	1	79	4	10.3	0.4000	648	1	75	1	7.1	0.2700
649	2	65	4	9.4	0.1020	650	1	79	4	4.1	0.0710	651	3	66	1	8.2	0.2890
652	2	77	3	7.1	0.7700	653	1	64	4	5.5	0.3060	654	1	65	4	3.0	0.0330
655	1	69	4	5.2	0.3030	656	1	82	4	7.1	0.0070	657	1	55	4	6.9	0.0600
658	1	61	4	4.8	0.4180	659	1	79	1	5.9	0.0510	660	3	64	1	6.0	0.1050
661	1	78	1	5.5	0.1480	662	1	64	4	4.7	0.1500	663	1	74	4	6.5	0.0840
664	1	75	4	5.2	0.1080	665	3	74	4	8.2	1.7740	666	1	66	4	7.6	0.9900
667	1	72	4	6.3	1.5070	668	1	63	4	6.4	0.0042	669	1	67	4	7.5	0.2830
670	1	62	3	6.2	0.1910	671	1	80	4	5.8	0.1180	672	1	63	4	6.6	0.0479
673	1	62	4	8.4	0.2160	674	3	68	1	5.9	0.3710	675	2	64	4	7.0	0.9500
676	1	64	4	6.3	0.2410	677	1	63	4	5.5	0.0042	678	1	60	4	4.6	0.0730
679	1	60	4	4.9	0.0800	680	1	70	4	7.0	0.2010	681	1	81	1	6.7	0.0600
682	1	73	4	7.2	0.2670												

Bibliography

Agresti, A.: 2002, *Categorical Data Analysis*, Wiley Series in Probability and Statistics, 2nd edition edn, Wiley.

Agresti, A., Booth, J. G., Hobert, J. P. and Caffo, B.: 2000, Random-effects modelling of categorical response data, *Sociological Methodology* **30**, 27–80.

Agresti, A. and Lang, J. B.: 1993, A proportional odds model with subject-specific effects for repeated ordered categorical responses, *Biometrika* **80**(3), 527–534.

Agresti, A. and Liu, I.: 2001, Strategies for modelling a categorical variable allowing multiple category choices, *Sociological Methods & Research* **29**(4), 403–434.

Agresti, A. and Liu, I.-M.: 1998, Modelling responses to a categorical variable allowing arbitrarily many category choices, *Technical Report Technical Report 575*, University of Florida, Department of Statistics.

Agresti, A. and Liu, I. M.: 1999, Modelling a categorical variable allowing arbitrarily many category choices, *Biometrics* **55**(3), 936–943.

Agresti, A. and Natarajan, R.: 2001, Modelling clustered ordered categorical data: A survey, *International Statistical Review* **69**(3), 345–371.

Aitchison, J.: 1962, Large sample restricted parametric tests, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **24**(1), 234–250.

Aitchison, J. and Silvey, S. D.: 1958, Maximum likelihood estimation of parameters subject to restraints, *Annals of Mathematical Statistics* **29**(3), 813–828.

Aitchison, J. and Silvey, S. D.: 1960, Maximum likelihood estimation procedures and associated tests of significance, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **22**(1), 154–171.

Ali, M. W. and Talukder, E.: 2005, Analysis of longitudinal binary data with missing data due to dropouts, *Journal of Biopharmaceutical Statistics* **15**(6), 993–1007.

Anderson, E. B.: 1980, *Discrete Statistical Models with Social Science Applications*, North-Holland, New York.

Anderson, J. and Philips, P.: 1981, Regression, discrimination and measurement models for ordered categorical variables, *Applied Statistics-Journal of the Royal Statistical Society Series C* **50**, 22–31.

Andrews, D.: 1994, Empirical process methods in econometrics, *in* R. Engle and D. McFadden (eds), *Handbook of Econometrics*, Vol. 4, Elsevier Science, pp. 2247–2294.

Arbogast, P. G. and Lin, D. Y.: 2005, Model-checking techniques for stratified case-control studies, *Statistics in Medicine* **24**(2), 229–247.

Atkinson, A. C. and Riani, M.: 1997, Bivariate boxplots, multiple outliers, multivariate transformations and discriminant analysis: The 1997 Hunter Lecture, *Environmetrics* **8**(6), 583–602.

Baade, I. A. and Pettitt, A. N.: 2000, Multiple and conditional deletion diagnostics for general linear models, *Communications in Statistics-Theory and Methods* **29**(8), 1899–1910.

Banerjee, M. and Frees, E. W.: 1997, Influence diagnostics for linear longitudinal models, *Journal of the American Statistical Association* **92**(439), 999–1005.

Bell, R., Rose, C. and Damon, A.: 1966, The veterans administration longitudinal study of healthy aging, *Gerontologist* **6**, 179–184.

Belsley, D., Kuh, E. and Welsch, R.: 1980, *Regression Diagnostics*, Wiley, New York.

Bennett, S.: 1983a, Analysis of survival data by the proportional odds model, *Statistics in Medicine* **2**, 273–277.

Bennett, S.: 1983b, Log-logistic regression-models for survival data, *Applied Statistics-Journal of the Royal Statistical Society Series C* **32**(2), 165–171.

Bergsma, W.: 1997, *Marginal Models for Categorical Data*, Tilburg University Press, Tilburg.

Bilder, C. R. and Loughin, T. M.: 2002, Testing for conditional multiple marginal independence, *Biometrics* **58**(1), 200–208.

Bilder, C. R. and Loughin, T. M.: 2004, Testing for marginal independence between two categorical variables with multiple responses, *Biometrics* **60**(1), 241–248.

Bilder, C. R., Loughin, T. M. and Nettleton, D.: 2000, Multiple marginal independence testing for pick any/c variables, *Communications in Statistics-Simulation and Computation* **29**(4), 1285–1316.

Birch, M. W.: 1965, The detection of partial association ii: The general case, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **27**(1), 111–124.

Bishop, Y., Fienberg, S. and Holland, P.: 1975, *Discrete Multivariate Analysis*, MIT Press, Cambridge, Mass.

Booth, J. G. and Hobert, J. P.: 1998, Standard errors of prediction in generalized linear mixed models, *Journal of the American Statistical Association* **93**(441), 262–272.

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.

Booth, J. G. and Sakar, S.: 1998, Monte carlo approximation of bootstrap variances, *The American Statistician* **52**(4), 354–357.

Brant, R.: 1990, Assessing proportionality in the proportional odds model for ordinal logistic regression, *Biometrics* **46**(4), 1171–1178.

Breslow, N.: 1981, Odds ratio estimators when the data are sparse, *Biometrika* **68**(1), 73–84.

Breslow, N. and Day, N.: 1980, *Statistical Methods in Cancer Research I: The Analysis of Case-Control Studies*, International Agency for Research on Cancer, Lyon.

Breslow, N. E.: 1984, Extra-poisson variation in log-linear models, *Applied*

Statistics-Journal of the Royal Statistical Society Series C **33**(1), 38–44.

Breslow, N. E. and Clayton, D. G.: 1993, Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**(421), 9–25.

Breslow, N. E. and Liang, K. Y.: 1982, The variance of the Mantel-Haenszel estimator, *Biometrics* **38**(4), 943–952.

Breslow, N. E. and Lin, X. H.: 1995, Bias correction in generalized linear mixed models with a single component of dispersion, *Biometrika* **82**(1), 81–91.

Carey, V., Zeger, S. L. and Diggle, P.: 1993, Modelling multivariate binary data with alternating logistic regressions, *Biometrika* **80**(3), 517–526.

Chatterjee, S. and Hadi, A.: 1986, Influential observations, high leverage points, and outliers in linear regression, *Statistical Science* **1**(3), 379–393.

Chatterjee, S. and Hadi, A.: 1988, *Sensitivity Analysis in Linear Regression*, Wiley series in probability and mathematical statistics, Wiley, New York.

Chen, M. H., Tong, X. W. and Sun, J. G.: 2007, The proportional odds model for multivariate interval-censored failure time data, *Statistics in Medicine* **26**(28), 5147–5161.

Christensen, R., Pearson, L. M. and Johnson, W.: 1992, Case-deletion diagnostics for mixed models, *Technometrics* **34**(1), 38–45.

Clayton, D. G.: 1974, Some odds ratio statistics for analysis of ordered categorical data, *Biometrika* **61**(3), 525–531.

Cochran, W. G.: 1954, Some methods for strengthening the common 2x2 tests, *Biometrics* **10**(4), 417–451.

Cook, R. D.: 1977, Detection of influential observation in linear regression, *Technometrics* **19**(1), 15–18.

Cook, R. D. and Weisberg, S.: 1982, *Residuals and Influence in Regression*, Chapman and Hall, London.

Cook, R. D. and Weisberg, S.: 1997, Graphics for assessing the adequacy of regression models, *Journal of the American Statistical Association* **92**(438), 490–499.

Coombs, C.: 1964, *A Theory of Data*, Wiley, New York.

Davidson, J.: 1994, *Stochastic Limit Theory: An Introduction for Econometricians*, Advanced texts in econometrics, Oxford University Press, Oxford.

Dempster, A. P., Laird, N. M. and Rubin, D. B.: 1977, Maximum likelihood from incomplete data via EM algorithm, *Journal of the Royal Statistical Society Series B-Methodological* **39**(1), 1–38.

Efron, B. and Tibshirani, R.: 1993, *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Emerson, J.: 1994, Combining estimates of the odds ratio: The state of the art, *Statistical Methods in Medical Research* **3**, 157–178.

Fahrmeir, L. and Tutz, G.: 2001, *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer series in statistics, 2nd edn, Springer, New York.

Fay, M. P.: 2002, Measuring a binary response's range of influence in logistic regression, *American Statistician* **56**(1), 5–9.

Fitzmaurice, G., Laird, N. and Ware, J.: 2004, *Applied Longitudinal Analysis*, Wiley, New Jersey.

Fitzmaurice, G. M. and Laird, N. M.: 1993, A likelihood-based method for analyzing longitudinal binary responses, *Biometrika* **80**(1), 141–151.

Fitzmaurice, G. M., Molenberghs, G. and Lipsitz, S. R.: 1995, Regression-models for longitudinal binary responses with informative drop-outs, *Journal of the Royal Statistical Society Series B-Methodological* **57**(4), 691–704.

Fung, W. K., Zhu, Z. Y., Wei, B. C. and He, X. M.: 2002, Influence diagnostics and outlier tests for semiparametric mixed models, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **64**, 565–579.

Gange, S. J.: 1995, Generating multivariate categorical variates using the iterative proportional fitting algorithm, *American Statistician* **49**(2), 134–138.

Gart, J. J.: 1962, On combination of relative risks, *Biometrics* **18**(4), 594–600.

Gelfand, A. E. and Carlin, B. P.: 1993, Maximum-likelihood-estimation for constrained-data or missing-data models, *Canadian Journal of Statistics-Revue Canadienne de Statistique* **21**(3), 303–311.

Geyer, C. J. and Thompson, E. A.: 1992, Constrained Monte-Carlo maximum-likelihood for dependent data, *Journal of the Royal Statistical Society Series B-Methodological* **54**(3), 657–699.

Glonek, G. F. V.: 1996, A class of regression models for multivariate categorical responses, *Biometrika* **83**(1), 15–28.

Glonek, G. F. V. and McCullagh, P.: 1995, Multivariate logistic models, *Journal of the Royal Statistical Society Series B-Methodological* **57**(3), 533–546.

Goldstein, H.: 1991, Nonlinear multilevel models, with an application to discrete response data, *Biometrika* **78**(1), 45–51.

Greenland, S.: 1989, Generalized Mantel-Haenszel estimators for K 2xJ tables, *Biometrics* **45**(1), 183–191.

Guilbaud, O.: 1983, On the large-sample distribution of the Mantel-Haenszel odds-ratio estimator, *Biometrics* **39**(2), 523–525.

Haber, M.: 1985, Maximum-likelihood methods for linear and log-linear models in categorical-data, *Computational Statistics & Data Analysis* **3**(1), 1–10.

Hampel, F. R.: 1974, Influence curve and its role in robust estimation, *Journal of the American Statistical Association* **69**(346), 383–393.

Hartzel, J., Agresti, A. and Caffo, B.: 2001, Multinomial logit random effects models, *Statistical Modelling* (1), 81–102.

Hartzel, J., Liu, I. M. and Agresti, A.: 2001, Describing heterogeneous effects in stratified ordinal contingency tables, with application to multi-center clinical trials, *Computational Statistics & Data Analysis* **35**(4), 429–449.

Harville, D.: 1976, Extension of Gauss-Markov theorem to include estimation of random effects, *Annals of Statistics* **4**(2), 384–395.

Haslett, J.: 1999, A simple derivation of deletion diagnostic results for the general linear model with correlated errors, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **61**, 603–609.

Haslett, J. and Dillane, D.: 2004, Application of 'delete = replace' to deletion diag-

nostics for variance component estimation in the linear mixed model, *Journal of the Royal Statistical Society Series B-Statistical Methodology* **66**, 131–143.

Haslett, J. and Haslett, S. J.: 2007, The three basic types of residuals for a linear model, *International Statistical Review* **75**(1), 1–24.

Hauck, W. W.: 1979, Large sample variance of the Mantel-Haenszel estimator of a common odds ratio, *Biometrics* **35**(4), 817–819.

Heagerty, P. J. and Zeger, S. L.: 1996, Marginal regression models for clustered ordinal measurements, *Journal of the American Statistical Association* **91**(435), 1024–1036.

Hedeker, D. and Gibbons, R.: 2006, *Longitudinal Data Analysis*, J. Wiley, Hoboken, NJ.

Hoaglin, D. C. and Welsch, R. E.: 1978, Hat matrix in regression and Anova, *American Statistician* **32**(1), 17–22.

Hosmer, D. and Lemeshow, S.: 2000, *Applied Logistic Regression*, 2nd edn, Wiley, New York.

Hunter, D. R. and Lange, K.: 2002, Computing estimates in the proportional odds model, *Annals of the Institute of Statistical Mathematics* **54**(1), 155–168.

Khmaladze, E. V. and Koul, H. L.: 2004, Martingale transforms goodness-of-fit tests in regression models, *Annals Of Statistics* **32**(3), 995–1034.

Kim, J. H.: 2003, Assessing practical significance of the proportional odds assumption, *Statistics & Probability Letters* **65**(3), 233–239.

Kirmani, S. and Gupta, R. C.: 2001, On the proportional odds model in survival analysis, *Annals of the Institute of Statistical Mathematics* **53**(2), 203–216.

Laird, N. M. and Ware, J. H.: 1982, Random-effects models for longitudinal data, *Biometrics* **38**(4), 963–974.

Landis, J. R., Heyman, E. R. and Koch, G. G.: 1978, Average partial association in 3-way contingency-tables - review and discussion of alternative tests, *International Statistical Review* **46**(3), 237–254.

Lang, J. B.: 1996, Maximum likelihood methods for a generalized class of log-linear models, *Annals of Statistics* **24**(2), 726–752.

Lang, J. B.: 2004, Multinomial-Poisson homogeneous models for contingency tables, *Annals of Statistics* **32**(1), 340–383.

Lang, J. B.: 2005, Homogeneous linear predictor models for contingency tables, *Journal of the American Statistical Association* **100**(469), 121–134.

Lang, J. B. and Agresti, A.: 1994, Simultaneously modelling joint and marginal distributions of multivariate categorical responses, *Journal of the American Statistical Association* **89**(426), 625–632.

Langford, I. H. and Lewis, T.: 1998, Outliers in multilevel data, *Journal of the Royal Statistical Society Series A-Statistics In Society* **161**, 121–153.

Lawal, B.: 2003, *Categorical Data Analysis With SAS and SPSS Applications*, 1st edn, Lawrence Erlbaum.

Lawrance, A. J.: 1995, Deletion influence and masking in regression, *Journal of the Royal Statistical Society Series B-Methodological* **57**(1), 181–189.

Lee, A. H. and Fung, W. K.: 1997, Confirmation of multiple outliers in general-

ized linear and nonlinear regressions, *Computational Statistics & Data Analysis* **25**(1), 55–65.

Lee, A. J.: 1993, Generating random binary deviates having fixed marginal distributions and specified degrees of association, *American Statistician* **47**(3), 209–215.

Lee, S. Y. and Lu, B.: 2003, Case-deletion diagnostics for nonlinear structural equation models, *Multivariate Behavioral Research* **38**(3), 375–400.

Lee, S. Y. and Xu, L. A.: 2003, Case-deletion diagnostics for factor analysis models with continuous and ordinal categorical data, *Sociological Methods & Research* **31**(3), 389–419.

Liang, K. Y.: 1987, Extended Mantel-Haenszel estimating procedure for multivariate logistic-regression models, *Biometrics* **43**(2), 289–299.

Liang, K. Y. and Self, S. G.: 1985, Tests for homogeneity of odds ratio when the data are sparse, *Biometrika* **72**(2), 353–358.

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

Liang, K. Y., Zeger, S. L. and Qaqish, B.: 1992, Multivariate regression-analyses for categorical data, *Journal of the Royal Statistical Society Series B-Methodological* **54**(1), 3–40.

Lin, D. Y. and Spiekerman, C. F.: 1996, Model checking techniques for parametric regression with censored data, *Scandinavian Journal of Statistics* **23**(2), 157–177.

Lin, D. Y. and Wei, L. J.: 1991, Goodness-of-fit tests for the general Cox regression-model, *Statistica Sinica* **1**(1), 1–17.

Lin, D. Y., Wei, L. J. and Ying, Z.: 1993, Checking the Cox model with cumulative sums of martingale-based residuals, *Biometrika* **80**(3), 557–572.

Lin, D. Y., Wei, L. J. and Ying, Z.: 2002, Model-checking techniques based on cumulative residuals, *Biometrics* **58**(1), 1–12.

Lipsitz, S. R., Fitzmaurice, G. M. and Molenberghs, G.: 1996, Goodness-of-fit tests for ordinal response regression models, *Applied Statistics-Journal of the Royal Statistical Society Series C* **45**(2), 175–190.

Lipsitz, S. R., Kim, K. and Zhao, L. P.: 1994, Analysis of repeated categorical-data using generalized estimating equations, *Statistics in Medicine* **13**(11), 1149–1163.

Lipsitz, S. R., Laird, N. M. and Harrington, D. P.: 1991, Generalized estimating equations for correlated binary data - using the odds ratio as a measure of association, *Biometrika* **78**(1), 153–160.

Little, R. J. A.: 1995, Modelling the drop-out mechanism in repeated-measures studies, *Journal of the American Statistical Association* **90**(431), 1112–1121.

Little, R. and Rubin, D.: 1987, *Statistical Analysis with Missing Data*, Wiley series in probability and mathematical statistics. Applied probability and statistics, John Wiley and Sons, New York.

Liu, I.: 1995, *Mantel-Haenszel-Type Inference for Odds Ratios with Ordinal Responses*, PhD thesis, University of Florida.

Liu, I.: 2003, Describing ordinal odds ratios for stratified r x c tables, *Biometrical Journal* **45**(6), 730–750.

Liu, I. and Agresti, A.: 2005, The analysis of ordered categorical data: An overview and a survey of recent developments, *Test* **14**(1), 1–30.

Liu, I. M. and Agresti, A.: 1996, Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response, *Biometrics* **52**(4), 1223–1234.

Liu, I., Mukherjee, B., Suesse, T., Sparrow, D. and Park, K. P.: 2008, Graphical diagnostics to check model misspecification for the proportional odds regression model, *Statistics in Medicine* pp. 18, in print.

Liu, I. and Suesse, T.: 2008, The analysis of stratified multiple responses, *Biometrical Journal* **50**(1), 135–149.

Liu, I. and Wang, D. Q.: 2007, Diagnostics for stratified clinical trials in proportional odds models, *Communications in Statistics-Theory and Methods* **36**(1-4), 211–220.

Liu, Q. and Pierce, D. A.: 1993, Heterogeneity in Mantel-Haenszel-type models, *Biometrika* **80**(3), 543–556.

Liu, Q. and Pierce, D. A.: 1994, A note on Gauss-Hermite quadrature, *Biometrika* **81**(3), 624–629.

Loughin, T. M. and Scherer, P.: 1998, Testing for association in contingency tables with multiple column responses, *Biometrics* **54**, 630–637.

Louis, T. A.: 1982, Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society Series B-Methodological* **44**(2), 226–233.

Lu, W. B. and Zhang, H. H.: 2007, Variable selection for proportional odds model, *Statistics in Medicine* **26**(20), 3771–3781.

Mantel, N.: 1963, Chi-square tests with 1 degree of freedom - extensions of Mantel-Haenszel procedure, *Journal of the American Statistical Association* **58**(303), 690–700.

Mantel, N.: 1978, Marginal homogeneity, symmetry, and independence, *Communications in Statistics Part A-Theory and Methods* **7**(10), 953–976.

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.

McCullagh, P.: 1980, Regression-models for ordinal data, *Journal of the Royal Statistical Society Series B-Methodological* **42**(2), 109–142.

McCullagh, P. and Nelder, J. A.: 1983, *Generalized Linear Models*, 1st edn, Chapman and Hall, New York.

McCullagh, P. and Nelder, J. A.: 1989, *Generalized Linear Models*, 2nd edn, Chapman and Hall, New York.

McCulloch, C. E.: 1997, Maximum likelihood algorithms for generalized linear mixed models, *Journal of the American Statistical Association* **92**(437), 162–170.

Mickey, R. M. and Elashoff, R. M.: 1985, A generalization of the Mantel-Haenszel estimator of partial association for 2xJxK-tables, *Biometrics* **41**(3), 623–635.

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**(4), 1033–1044.

Munoz-Pichardo, J., Munoz-Garcia, J., Moreno-Rebollo, J. and Pino-Mejias, R.: 1995, A new approach to influence analysis in linear models, *Sankhya* **57**, 393–409.

Murphy, S. A., Rossini, A. J. and vanderVaart, A. W.: 1997, Maximum likelihood estimation in the proportional odds model, *Journal of the American Statistical Association* **92**(439), 968–976.

Nakanish, S., Yamane, K., Kamei, N., Okubo, M. and Kohno, N.: 2003, Elevated C-Reactive protein is a risk factor for the development of type-2 diabetes in Japanese Americans, *Diabetes Care* **26**, 2754–2757.

Nelder, J. A. and Wedderburn, R.: 1972, Generalized linear models, *Journal of the Royal Statistical Society Series A-General* **135**(3), 370–384.

Neter, J., Wasserman, W. and Kutner, M.: 1985, *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs.*, R.D. Irwin, Homewood, Illinois.

Pan, W.: 2002, Goodness-of-fit tests for GEE with correlated binary data, *Scandinavian Journal Of Statistics* **29**(1), 101–110.

Pan, Z. Y. and Lin, D. Y.: 2005, Goodness-of-fit methods for generalized linear mixed models, *Biometrics* **61**(4), 1000–1009.

Pardoe, I. and Cook, R. D.: 2002, A graphical method for assessing the fit of a logistic regression model, *American Statistician* **56**(4), 263–272.

Paul, S. R. and Donner, A.: 1989, A comparison of tests of homogeneity of odds ratios in K 2x2 tables, *Statistics in Medicine* **8**(12), 1455–1468.

Peixoto, J. L. and Lamotte, L. R.: 1989, Simultaneous identification of outliers and predictors using variable selection techniques, *Journal of Statistical Planning and Inference* **23**(3), 327–343.

Pena, D. and Yohai, V. J.: 1995, The detection of influential subsets in linear-regression by using an influence matrix, *Journal of the Royal Statistical Society Series B-Methodological* **57**(3), 611–611.

Perevozskaya, I., Rosenberger, W. F. and Haines, L. M.: 2003, Optimal design for the proportional odds model, *Canadian Journal of Statistics-Revue Canadienne de Statistique* **31**(2), 225–235.

Peterson, B. and Harrell, F. E.: 1990, Partial proportional odds models for ordinal response variables, *Applied Statistics-Journal of the Royal Statistical Society Series C* **39**(2), 205–217.

Pettitt, A. N.: 1984, Proportional odds models for survival-data and estimates using ranks, *Applied Statistics-Journal of the Royal Statistical Society Series C* **33**(2), 169–175.

Plackett, R. L.: 1981, *The Analysis of Categorical Data*, 2nd edn, Charles Griffin House, London.

Pliquett, R., Fasshauer, M., Blueher, M. and Paschke, R.: 2004, Neurohumoral stimulation in type-2-diabetes as an emerging disease concept., *Cardiovascular Diabetology* **3**.

Pregibon, D.: 1981, Logistic regression diagnostics, *Annals of Statistics* **9**(4), 705–

724.

Preisser, J. S. and Perin, J.: 2007, Deletion diagnostics for marginal mean and correlation model parameters in estimating equations, *Statistics and Computing* **17**(4), 381–393.

Preisser, J. S. and Qaqish, B. F.: 1996, Deletion diagnostics for generalised estimating equations, *Biometrika* **83**(3), 551–562.

Prentice, R. L.: 1988, Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**(4), 1033–1048.

Prentice, R. L. and Zhao, L. P.: 1991, Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics* **47**(3), 825–839.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A.: 2002, Reliable estimation of generalized linear mixed models using adaptive quadrature, *The Stata Journal* **2**(1), 1–21.

Rasch, G.: 1961, On the general laws of and the meaning of measurement in psychology, *4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, pp. 321–355.

Raudenbush, S. W., Yang, M. L. and Yosef, M.: 2000, Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation, *Journal of Computational and Graphical Statistics* **9**(1), 141–157.

Richert, B. T., Tokach, M. D., Goodband, R. D. and Nelssen, J. L.: 1993, Integrated Swine Systems: 'The Animal Component' -Phase 1; The Kansas State University Survey, Swine Day 1993, *Technical report*, Kansas State University.

Robins, J., Breslow, N. and Greenland, S.: 1986, Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models, *Biometrics* **42**(2), 311–323.

Robins, J. M., Rotnitzky, A. and Zhao, L. P.: 1995, Analysis of semiparametric regression-models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association* **90**(429), 106–121.

Sato, T.: 1991, An estimating equation approach for the analysis of case-control studies with exposure measured at several levels, *Statistics in Medicine* **10**(7), 1037–1042.

Schall, R.: 1991, Estimation in generalized linear-models with random effects, *Biometrika* **78**(4), 719–727.

Searle, S.: 1982, *Matrix Algebra Useful for Statistics*, Wiley, New York.

Sen, P. K. and Singer, J. M.: 1993, *Large Sample Methods in Statistics: An Introduction with Applications*, Chapman & Hall, New York.

Shao, J.: 1999, *Mathematical Statistics*, Springer Series in Statistics, Springer, New York.

Shao, J. and Tu, D.: 1995, *The Jackknife and Bootstrap*, Springer Series in Statistics, Springer, New York.

Silvey, S. D.: 1959, The Lagrangian multiplier test, *Annals of Mathematical Statistics* **30**(2), 389–407.

Simon, G.: 1974, Alternative analyses for singly-ordered contingency table, *Journal of the American Statistical Association* **69**(348), 971–976.

Simonoff, J. S. and Tsai, C. L.: 1991, Assessing the influence of individual observations on a goodness-of-fit test based on nonparametric regression, *Statistics & Probability Letters* **12**(1), 9–17.

Spiekerman, C. F. and Lin, D. Y.: 1996, Checking the marginal Cox model for correlated failure time data, *Biometrika* **83**(1), 143–156.

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.

Stiratelli, R., Laird, N. and Ware, J. H.: 1984, Random-effects models for serial observations with binary response, *Biometrics* **40**(4), 961–971.

Su, J. Q. and Wei, L. J.: 1991, A lack-of-fit test for the mean function in a generalized linear model, *Journal of the American Statistical Association* **86**(414), 420–426.

Suesse, T. and Liu, I.: 2008, Diagnostics for multiple response data, *Proceedings of PROBASTAT 2006*, Vol. 39, Tatra Mountains Publications, Smolenice Castle, Slovak Republic, pp. 105–113.

Sun, J. G., Sun, L. Q. and Zhu, C.: 2007, Testing the proportional odds model for interval-censored data, *Lifetime Data Analysis* **13**(1), 37–50.

Sundaram, R.: 2006, Semiparametric inference for the proportional odds model with time-dependent covariates, *Journal of Statistical Planning And Inference* **136**(2), 320–334.

Tarone, R. E.: 1985, On heterogeneity tests based on efficient scores, *Biometrika* **72**(1), 91–95.

Team R Development Core, A Language and Environment for Statistical Computing: 2006.

The Expert committee on the diagnosis and classification of diabetes, Report of the Expert committee on the diagnosis and classification of diabetes mellitus: 1997.

Toledano, A. Y. and Gatsonis, C.: 1996, Ordinal regression methodology for ROC curves derived from correlated data, *Statistics in Medicine* **15**(16), 1807–1826.

Tutz, G. and Hennevogl, W.: 1996, Random effects in ordinal regression models, *Computational Statistics & Data Analysis* **22**(5), 537–557.

Vaart, A. W. and Wellner, J. A.: 1996, *Weak Convergence and Empirical Processes*, Springer series in statistics, Springer, New York.

Wang, D. Q., Critchley, F. and Liu, I.: 2004, Diagnostics analysis and perturbations in a clustered sampling model, *Communications in Statistics-Theory and Methods* **33**(11-12), 2709–2721.

Wang, H. M., Jones, M. P. and Storer, B. E.: 2006, Comparison of case-deletion diagnostic methods for Cox regression, *Statistics in Medicine* **25**(4), 669–683.

Wedderburn, R.: 1974, Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61**, 439–447.

Williams, D. A.: 1982, Extra-binomial variation in logistic linear-models, *Applied Statistics-Journal of the Royal Statistical Society Series C* **31**(2), 144–148.

Williams, D. A.: 1987, Generalized linear-model diagnostics using the deviance and single case deletions, *Applied Statistics-Journal of the Royal Statistical Society Series C* **36**(2), 181–191.

Williams, O. D. and Grizzle, J. E.: 1972, Analysis of contingency tables having ordered response categories, *Journal of the American Statistical Association* **67**(337), 55–63.

Wilson, E.: 1927, Probable inference, the law of succession, and statistical inference, *Journal of the American Statistical Association* **22**, 209–212.

Wu, C. O.: 1995, Estimating the real parameter in a 2-sample proportional odds model, *Annals of Statistics* **23**(2), 376–395.

Xiang, L. M., Tse, S. K. and Lee, A. H.: 2002, Influence diagnostics for generalized linear mixed models: applications to clustered data, *Computational Statistics & Data Analysis* **40**(4), 759–774.

Yan, J.: 2004, geepack: Yet another package for generalized estimating equations.

Yan, J. and Fine, J.: 2004, Estimating equations for association structures, *Statistics in Medicine* **23**(6), 859–874.

Yanagawa, T. and Fujii, Y.: 1990, Homogeneity test with a generalized Mantel-Haenszel estimator for L 2 x K contingency-tables, *Journal of the American Statistical Association* **85**(411), 744–748.

Yanagawa, T. and Fujii, Y.: 1995, Projection-method Mantel-Haenszel estimator for K 2xJ tables, *Journal of the American Statistical Association* **90**(430), 649–656.

Yang, S. and Prentice, R. L.: 1999, Semiparametric inference in the proportional odds regression model, *Journal of the American Statistical Association* **94**(445), 125–136.

Zeger, S. L., Liang, K. Y. and Albert, P. S.: 1988, Models for longitudinal data - a generalized estimating equation approach, *Biometrics* **44**(4), 1049–1060.

Zeng, D. L., Lin, D. Y. and Yin, G. S.: 2005, Maximum likelihood estimation for the proportional odds model with random effects, *Journal of the American Statistical Association* **100**(470), 470–483.

Zewotir, T. and Galpin, J. S.: 2006, Evaluation of linear mixed model case deletion diagnostic tools by Monte Carlo simulation, *Communications in Statistics-Simulation and Computation* **35**(3), 645–682.

Zhang, J. and Boos, D.: 1996, Generalized Cochran-Mantel-Haenszel test statistics for correlated categorical data, *Technical report*, Department of Statistics, North Carolina State University.

Zhao, L. P. and Prentice, R. L.: 1990, Correlated binary regression using a quadratic exponential model, *Biometrika* **77**(3), 642–648.

Ziegler, A., Blettner, M., Kastner, C. and Chang-Claude, J.: 1998, Identifying influential families using regression diagnostics for generalized estimating equations, *Genetic Epidemiology* **15**(4), 341–353.

Symbols/Notations

Symbol	Explanation
$\mathbb{R}$, $\mathbb{R}^s$, $\mathbb{R}^{s\times t}$	space of real numbers with dimensions: 1, s and $s \times t$
$a \in \mathbb{R}$	scalar
$\mathbf{A} \in \mathbb{R}^{s\times t}$	$s \times t$ real valued matrix
$\mathbf{a} \in \mathbb{R}^s$	s dimensional real valued column vector
$\mathbf{A}^T$, $\mathbf{a}^T$	transpose of matrix $\mathbf{A}$, transpose of vector $\mathbf{a}$
$\|\mathbf{a}\|_p$, $\|a\|$	p-norm of vector $\mathbf{a}$, absolute value of scalar a
$\mathbf{I}_s$	identity matrix of size $s \times s$
$\mathbf{1}_s$	column vector containing only ones of size s
Diag$(a_1, \ldots, a_n)$	diagonal matrix with elements $a_1, \ldots, a_n$ on diagonal
Diag$(\mathbf{A}_1, \ldots, \mathbf{A}_n)$	block-diagonal matrix with matrices $\mathbf{A}_1, \ldots, \mathbf{A}_n$ on diagonal
$\otimes$	Kronecker operator
$\mathbb{1}_{\{exp\}}$ or $\mathbb{1}(exp)$	indicator function, is one if expression exp is true and zero otherwise
$\rightarrow_d$	convergence in distribution
$\rightarrow_d$	convergence in probability
E, Var, Cov	expectation, variance and covariance
E^a, Var^a, Cov^a	asymptotic expectation, variance and covariance
$\widehat{par}$, $\hat{par}$	indicates estimator for parameter par
$\frac{\partial f}{\partial y}$	partial derivative of function f with respect to y
$\chi^2(k)$, $df = k$	chi-squared distribution with k degree of freedom (df)