Modelling Strategies for Repeated Multiple Response Data

Thomas Suesse
School of Mathematics and Applied Statistics
University of Wollongong, Australia
Ivy Liu ('responsible author')
School of Mathematics, Statistics and Operations Research
Victoria University of Wellington, New Zealand

June 13, 2010

Abstract

Agresti & Liu (2001) discussed modelling strategies for a multiple response variable, a categorical variable for which respondents can select any number of outcome categories. This article discusses modelling strategies of a repeated multiple response variable, a categorical variable for which respondents can select any number of categories on repeated occasions. We consider each of the responses as a binary response and model the mean binary responses with two approaches: A marginal model approach and a mixed model approach. For the marginal model approach, we propose another more efficient estimation method based on groupwise correlation structures applying generalised estimating equations (GEE). A simulation study evaluates the performance of the new groupwise method. We illustrate different approaches using an example.

Keywords: Multiple Responses, Repeated Measurements, GLM, GEE, GLMM

1 Introduction

Surveys often contain qualitative variables for which respondents may select any number out of c outcome categories. The respondents are asked to "tick all that apply" on a list of the outcome categories. Categorical variables that summarise this type of data are called *multiple response variables*. As an example, for the question "When you go out to bars, do you go out to...?" with possible responses (a) Socialise with friends, (b) Meet new people, (c) Listen to music, (d) Get drunk, (e) Other (please specify), respondents would be introduced to select whichever of the outcomes apply. Each outcome category is referred to as an *item* (Agresti & Liu 1999).

Multiple responses have been considered in the literature by various authors. For instance, Loughin & Scherer (1998) developed a large–sample weighted chi-squared test and a small–sample bootstrap test for the independence between each of the c items and an explanatory variable. Agresti & Liu (1999, 2001) discussed different modelling strategies to describe the association between items and explanatory variables. When the data are stratified by a third variable, Bilder & Loughin (2002) provided a test for the conditional multiple marginal independence to detect whether the group and items are marginally independent given the stratification variable. Furthermore, Bilder & Loughin (2004) gave a test for marginal independence between two categorical variables with multiple responses. Besides the modelling strategies and testing methods, Liu & Suesse (2008) presented two methods (GEE and Mantel-Haenszel) to make inferences across multiple responses when data include highly stratified variables. None of the above papers considered the situations in a longitudinal manner where respondents were asked on several occasions. This paper discusses modelling strategies generalised from the ones introduced by Agresti & Liu (2001) to analyse the repeated multiple response data when the multiple response data were collected at several time points.

Agresti & Liu (2001) treated the responses for each of the items as binary responses (being selected or not). Then, they modelled these correlated responses using the marginal model approach and the mixed model approach. For the repeated multiple responses, we also treat the responses for each of the items as binary responses. Yet, these binary responses are correlated in two levels, across both items and different time points.

Section 2 considers the marginal model approach for repeated multiple response data. We primarily focus on generalised estimating equations and consider a variety of possible correlation structures. We also propose a more efficient groupwise method assuming different group correlation parameters in Section 3. Section 4 gives several feasible fitting techniques for the mixed model approach. In Section 5, we conduct a simulation study to investigate the performance of the new groupwise method. Section 6 illustrates methods for a survey data set about the choice of bars based on their features. The final section finishes with a discussion.

2 Marginal Modelling

Let $y_{ijt} = 1$ if subject $i = 1, \ldots, n$ selects category $j = 1, \ldots, c$ at time point or occasion $t = 1, \ldots, T$ and $y_{ijt} = 0$ otherwise. Let $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \ldots, \mathbf{y}_{iT}^T)^T$ denote the ith subject's $2^{c \cdot T}$ response profile for c items and T time points, where $\mathbf{y}_{it} = (y_{i1t}, y_{i2t}, \ldots, y_{ict})^T$. Note that superscript T denotes the transpose of a vector/matrix and subscript T refers to the number of time points. Denote the mean of y_{ijt} by $\pi_{j|it}$, the probability of a positive response on item j at occasion t by the ith subject. Define similarly the mean of $\mathbf{y}_i$ as $\boldsymbol{\pi}_i$. Let $\mathbf{X}_i = (\mathbf{X}_{i0}, \mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT})$ be a vector of covariates of the ith subject, where $\mathbf{X}_{it}, t = 1, \ldots, T$ are the time–variant covariates and $\mathbf{X}_{i0}$ the time–invariant covariates.

When $T = 1$, Agresti & Liu (2001) modelled the mean $\boldsymbol{\pi}_i$ of $\mathbf{y}_i$ in terms of the covariates $\mathbf{X}_i$. For the general case $T \geq 1$, we model the mean response $\pi_{j|it}$ in terms of the covariates $\mathbf{X}_{it}, t \geq 0$ by

$$h(\pi_{j|it}) = \alpha_{jt} + \mathbf{X}_{i0}\boldsymbol{\beta}_{0j} + \mathbf{X}_{it}\boldsymbol{\beta}_{jt} = \mathbf{Z}_{ijt}\boldsymbol{\beta}, \tag{1}$$

where $h(\cdot)$ is the link function, α_{jt} is the j-th intercept parameter at time t, $\mathbf{Z}_{ijt}$ is the corresponding design matrix depending on $\mathbf{X}_i$, and $\boldsymbol{\beta} := (\alpha_{11}, \ldots, \alpha_{cT}, \boldsymbol{\beta}_{01}^T, \ldots, \boldsymbol{\beta}_{cT}^T)^T$ the vector containing all parameters. For a special case of using the logit link, the model becomes

$$\log\left(\frac{\pi_{j|it}}{1-\pi_{j|it}}\right) = \alpha_{jt} + \mathbf{X}_{i0}\boldsymbol{\beta}_{0j} + \mathbf{X}_{it}\boldsymbol{\beta}_{jt}.$$

There are several fitting techniques for model (1). Naively, one can assume independence between all items and occasions and then use ordinary software for generalised linear models (McCullagh & Nelder 1989). However, assuming independence does not give proper standard error estimates for the parameter estimators.

Alternatively, model (1) can be expressed as a generalised log-linear model and the maximum likelihood (ML) method (Lang & Agresti 1994, Lang 1996) can be used to yield parameter estimates for $\boldsymbol{\beta}$. Agresti & Liu (1999) showed the ML approach for the special case of $T = 1$. An extension of the generalised log-linear model given by Lang (2005) allows any smooth link function. The ML method treats counts from the $2^{c \cdot T}$ response profile for each different covariate setting as a multinomial distribution. It maximises the multinomial likelihood subject to constraints satisfying the model. Because the number $2^{c \cdot T}$ is usually very large, the number of observations for many of the $2^{c \cdot T}$ categories will be very small (e.g. zero). This sparseness is even worse when some covariates are continuous. It causes problems on the ML fitting algorithm. The ML approach is plausible only when the number of subjects is large, $2^{c \cdot T}$ is small and all covariates are categorical with few levels.

Besides the ML approach, Agresti & Liu (1999) showed another fitting procedure using the generalised estimation equations (GEE) (Liang & Zeger 1986). The GEE method fits marginal models simultaneously and incorporates a chosen correlation structure. It is an extension of the quasi-likelihood method (Wedderburn 1974) for multivariate data. Denote $\text{Var}(\mathbf{y}_i) = \mathbf{f}_i \cdot \phi^{-1}$ with variance function $\mathbf{f}_i = \mathbf{f}(\boldsymbol{\pi}_i)$ $(= \boldsymbol{\pi}_i(\mathbf{1}_{cT} - \boldsymbol{\pi}_i)$ for binary responses, where $\mathbf{1}_{cT}$ is a vector ones of length cT), and the scale or dispersion parameter by ϕ. Suppose model (1) is true, then the GEE estimates are obtained by computing the root of the generalised estimation equations

$$\sum_{i=1}^{n} \mathbf{U}_i = 0 \text{ with } \mathbf{U}_i = \mathbf{M}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i,$$

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 $\mathbf{r}_i = (\mathbf{y}_i - \boldsymbol{\pi}_i)$. The dimension of matrix $\mathbf{M}_i$ is $cT \times p$, where p is the number of parameters in $\boldsymbol{\beta}$. Matrix $\mathbf{A}_i = \sqrt{\mathbf{f}_i}$ is diagonal of size $cT \times cT$ and $\mathbf{R}_i(\boldsymbol{\alpha})$ is the $cT \times cT$ correlation matrix for subject i $(i = 1, \ldots, n)$ depending on correlation parameter(s) $\boldsymbol{\alpha}$. The correlation matrix is based on a "working guess" about the correlation structure of the items across different occasions. Preisser & Qaqish (1996) suggested the iterated weighted least squares method to obtain the vector of estimators $\hat{\boldsymbol{\beta}}$. One can adjust the standard error of the parameter estimators to reflect what actually occurs for the sample data using a "sandwich" method.

We consider specific choices of the correlation structure $\mathbf{R}_i(\boldsymbol{\alpha})$ for multiple response data and repeated multiple response data. Choosing a good correlation structure is essential to obtain good variance estimates and more efficient parameter estimates for $\hat{\boldsymbol{\beta}}$, e.g. see simulation study in Liang & Zeger (1986).

Denote the correlation structure $\mathbf{R}_i = \text{Corr}(\mathbf{y}_i) = \mathbf{R}$ by

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \cdots & \mathbf{R}_{1T} \\ \mathbf{R}_{12} & \mathbf{R}_{22} & \cdots & \mathbf{R}_{2T} \\ \vdots & \vdots & & \vdots \\ \mathbf{R}_{1T} & \mathbf{R}_{2T} & \cdots & \mathbf{R}_{TT} \end{pmatrix},$$

where the indices t_1 and t_2 of $\mathbf{R}_{t_1t_2} \in \mathbb{R}^{c\times c}$ refer to two different occasions. We assume that all subjects have the same correlation structure. The sub-matrix $\mathbf{R}_{tt}$ refers to the correlation structure across different items at a given occasion t and the sub-matrix $\mathbf{R}_{t_1t_2}$ refers to the correlation structure across different items at a pair of occasions t_1 and t_2. They have the form:

$$\mathbf{R}_{tt} = \begin{pmatrix} 1 & R_{tt,12} & \cdots & R_{tt,1c} \\ R_{tt,12} & 1 & \cdots & R_{tt,2c} \\ \vdots & \vdots & & \vdots \\ R_{tt,1c} & R_{tt,2c} & \cdots & 1 \end{pmatrix}$$

$$\mathbf{R}_{t_1t_2} = \begin{pmatrix} R_{t_1t_2,11} & R_{t_1t_2,12} & \cdots & R_{t_1t_2,1c} \\ R_{t_1t_2,21} & R_{t_1t_2,22} & \cdots & R_{t_1t_2,2c} \\ \vdots & \vdots & & \vdots \\ R_{t_1t_2,c1} & R_{t_1t_2,c2} & \cdots & R_{t_1t_2,cc} \end{pmatrix}$$

Note, generally matrix $\mathbf{R}_{t_1t_2}$ is not symmetric, but $\mathbf{R}_{tt}$ is.

When $T = 1$, the matrix $\mathbf{R}$ reduces to $\mathbf{R}_{11}$ and we omit index t referring to the occasions. The most common correlation structures between items include

- independence (items): $R_{j_1j_2} = 0$ for all $j_1 \neq j_2$ (0 parameter)
- exchangeable (items): $R_{j_1j_2} = \alpha$ for all $j_1 \neq j_2$ (1 parameter)
- unstructured (items): totally unspecified $R_{i,j_1j_2} = \alpha_{j_1,j_2}$ ($\frac{1}{2}c(c-1)$ parameters)

and we estimate the parameters by the method of moments.

For repeated multiple responses ($T > 1$), we can assign the same correlation structure to $\mathbf{R}$ as the one mentioned above. In such a way, we can not distinguish between occasions and items. It is more appropriate to consider different structures for the submatrices of $\mathbf{R}_{tt}$ and $\mathbf{R}_{t_1t_2}$. Every submatrix $\mathbf{R}_{tt}$ can have the structures independence, exchangeable, and unstructured, as we considered for the special case of $T = 1$. For the off-diagonal matrices $\mathbf{R}_{t_1t_2}(\equiv \mathbf{R}_{t_2t_1})$ with $t_1 \neq t_2$, the diagonal elements do not usually equal one ($R_{t_1t_2,jj} \neq 1$) and the symmetry may not hold, i.e., $R_{t_1t_2,j_1j_2} \neq R_{t_1t_2,j_2j_1}$ for $j_1 \neq j_2$ and $t_1 \neq t_2$. Thus, the structure of $\mathbf{R}_{tt}$ does not apply. Instead, we define the following correlation structures for $\mathbf{R}_{t_1t_2}$:

- independence (items): $R_{it_1t_2,j_1j_2} = 0 \ \forall j_1, j_2 = 1, \dots, c$ (0 parameter)
- exchangeable (items): $R_{it_1t_2,j_1j_2} = \alpha \ \forall j_1, j_2 = 1, \dots, c$ (1 parameter)

- unstructured (items): totally unspecified (c^2 parameters).

This paper considers a common structure of $\mathbf{R}_{tt}$ and $\mathbf{R}_{t_1t_2}$ for all t and $t_1 \neq t_2$. One might assume different structures for different t. See Suesse (2008) for various choices.

For given time points, the correlation structure from the above options focuses on the dependence between items j_1 and j_2. We denote such a structure by "structure (items)". Alternatively, the correlation structure can be chosen based on the the dependence over time for given items j_1 and j_2. We denote such a structure by "structure (time)". The options include:

- exchangeable (time): $R_{t_1t_2,j_1j_2} = \alpha_{j_1j_2}$ (1 parameter)
- autoregressive (time): $R_{t_1t_2,j_1j_2} = \alpha_{j_1j_2}^{|t_1-t_2|}$ (1 parameter)
- unstructured (time): $R_{t_1t_2,j_1j_2} = \alpha_{t_1t_2,j_1j_2}$ ($T(T+1)/2$ parameters for $j_1 \neq j_2$ and $T(T-1)/2$ parameters for $j_1 = j_2$).

The option of exchangeable (time) and exchangeable (items) is equivalent to exchangeable for the whole matrix $\mathbf{R}$. Similarly, the option of unstructured (time) and unstructured (items) is equivalent to assuming unstructured for the whole matrix $\mathbf{R}$.

Combining the structures by assuming conditional structures for items given time points and for time points given items, seems a better approach than considering the structures $\mathbf{R}_{tt}$ and $\mathbf{R}_{t_1t_2}$ separately. In particular, the higher the number of time points is, the more plausible it is to take the time dependence structure into account.

3 Groupwise Correlation Estimation for GEE

Liang & Zeger (1986) assumed the correlation structure to be equal for all subjects and estimated the correlation parameters $\boldsymbol{\alpha}$ using the method of moments. This assumption is practical in terms of simplicity, but unrealistic. We assume the correlation model that allows the correlation parameters $\boldsymbol{\alpha}$ to vary for different groups (e.g., age, sex, etc). Suppose that there are a finite number of groups. Let the number of subjects in group g ($g = 1, \ldots, G$) be n_g with $\sum_{g=1}^{G} n_g = n$. Assume $\lim_{n\to\infty} n_g/n = a_g > 0$. The correlation parameters for group g are denoted by $\boldsymbol{\alpha}_g$. The GEE estimators have the following property:

Theorem 3.1 ("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

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

Liang & Zeger (1986) showed the proof for a special case of Theorem 3.1, where all subjects have the same correlation structure. Appendix A gives details of the proof for Theorem 3.1.

We label the groupwise correlation estimation method as *groupwise method* and the method assuming the same correlation parameters for all subjects as *standard method.* For the groupwise method, we require α_g to be $n_g^{1/2}$ consistent. In other words, we require n_g to be reasonable large. In Section 5, we conduct a simulation study to evaluate whether the groupwise method leads to more efficient estimators for $\boldsymbol{\beta}$.

4 Generalised Linear Mixed Models

The marginal model (1) is called a population–averaged model, which focuses on the marginal distribution of responses. Instead of assuming a particular distribution of responses, the GEE method specifies only the first two moments. The mean is linked to the predictor and the working correlation is incorporated to obtain the estimators. In contrast, generalised linear mixed models (GLMM) additionally include a subject specific effect, the random effect. This model is referred to as subject–specific modelling, since the model parameters are defined at the subject level.

Let $\mathbf{u}_i$ be the random effect vector for subject i and let $\mathbf{Q}_i$ be the design matrix for 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 $\mu_{ijt} = \mathbb{E}(y_{ijt}|\mathbf{u}_i)$. Given $\mathbf{u}_i$, the responses are independent on the same subject, which is known as the local independence assumption. Also, the responses are independent for different subjects. In our case, the distribution of y_{ijt} is binary and $\mu_{ijt} \equiv \pi_{j|it}$. The linear predictor for a GLMM is

$$h(\pi_{j|it}) = \mathbf{Z}_{ijt}\boldsymbol{\beta} + \mathbf{Q}_{ijt}\mathbf{u}_i,$$

where $\mathbf{Z}_{ijt}$, $\boldsymbol{h}(\cdot)$ and $\boldsymbol{\beta}$ have the same meaning as in model (1). The random effects $\mathbf{u}_i$ of dimension r ($r \leq c \times T$) are assumed to be multivariate 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 local independence assumption, the conditional density of $\mathbf{y}$ given $\mathbf{u}$ has the form

$$f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta}) \text{ with } f(\mathbf{y}_i|\mathbf{u}_i;\boldsymbol{\beta}) = \prod_{j=1}^{c}\prod_{t=1}^{T} f(y_{ijt}|\mathbf{u}_i;\boldsymbol{\beta}). \tag{2}$$

We can also write

$$f(\mathbf{u};\boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(\mathbf{u}_i;\boldsymbol{\Sigma}), \tag{3}$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ and $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$.

We maximise the likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y})$

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma}) d\mathbf{u} \tag{4}$$

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).

The integral usually cannot be solved analytically and numerical methods must be used. Gauss-Hermite quadrature methods directly approximate the integral (4). They work well for small dimension r of the random effect distribution, but become infeasible for a large r, because the number of quadrature points used to approximate the integral increases exponentially with r.

Some methods using the approximate likelihood are available (Stiratelli, Laird & Ware 1984, Schall 1991, Breslow & Clayton 1993, Zeger, Liang & Albert 1988, Goldstein 1991). However, most of them can yield poor estimates, in particular for first order expansions (Breslow & Lin 1995). Raudenbush, Yang & Yosef (2000) introduced a fast method combining a fully multivariate Taylor series expansion and a Laplace approximation, yielding accurate results. Other possible approaches include penalised log-likelihood equations (Schall 1991, Breslow & Clayton 1993), Bayesian mixed models (Fahrmeir & Tutz 2001) and Semi- or Nonparametric ML (Hartzel, Agresti & Caffo 2001).

This paper considers the indirect maximisation with the EM (expectation-maximisation) algorithm. One can treat the random effects $\mathbf{u}$ as unobserved data. Let $\boldsymbol{\Psi} = (\boldsymbol{\beta}^T, \boldsymbol{\Sigma}^T)^T$ that represents both the model parameters $\boldsymbol{\beta}$ and the parameters of the covariance matrix $\boldsymbol{\Sigma}$. Suppose that both $\mathbf{y}$ and $\mathbf{u}$ are observed, the complete likelihood becomes $f(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma})$. The log-likelihood is (McCulloch 1997)

$$\log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\beta}) + \log f(\mathbf{u}_i; \boldsymbol{\Sigma}). \tag{5}$$

The EM algorithm has two steps. Define

$$Q^{(0)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}') = \mathbb{E}(\log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\Psi})|\mathbf{y}; \boldsymbol{\Psi}') = \int \log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\Psi}) f(\mathbf{u}|\mathbf{y}; \boldsymbol{\Psi}') \mathbf{du},$$

where $\boldsymbol{\Psi}'$ is an old estimate in an iteration scheme and $\boldsymbol{\Psi}$ is the new estimate. First we compute the expectation in $Q^{(0)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}')$ (E-step) and then we maximise this expression (M-step) with respect to $\boldsymbol{\Psi}$ for given $\boldsymbol{\Psi}'$. Since the first term of the log-likelihood (5) depends on $\boldsymbol{\beta}$ and the second on $\boldsymbol{\Sigma}$, the form is equivalent to

$$Q^{(0)}(\boldsymbol{\Psi}|\boldsymbol{\Psi}') = \mathbb{E}(\log f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta})|\mathbf{y}; \boldsymbol{\Psi}') + \mathbb{E}(\log f(\mathbf{u}; \boldsymbol{\Sigma})|\mathbf{y}; \boldsymbol{\Psi}').$$

Therefore the M-step and E-step can be performed separately for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

McCulloch (1997) proposed several algorithms for the maximisation of the marginal likelihood using the EM algorithm. For our model fitting purposes, we use a combination of his Monte-Carlo-Newton-Raphson and simulated maximum likelihood algorithms. To approximate the integral numerically by Monte-Carlo approximation, we sample from $f(\mathbf{u}|\mathbf{y}; \boldsymbol{\Psi}')$ using the algorithm given by Booth & Hobert (1999) to control the Monte-Carlo error.

5 Simulation Study

In this section, we conduct a simulation study to investigate the performance of the groupwise correlation for the GEE method. Consider the model

$$\log\left(\frac{\pi_{j|ig}}{1-\pi_{j|ig}}\right) = X_{igj}\beta_j,\ i=1,...,n_g,\ g=1,\ldots,G,\ j=1,...,c, \tag{6}$$

where G is the number of groups and c is the number of items. For simplicity, we consider only one time point and choose $G=4$, $c=3$. The correlation structure for the gth group has the following form

$$\mathbf{R}_{ig} = \begin{pmatrix} 1 & R_{ig,12} & R_{ig,13} \\ R_{ig,12} & 1 & R_{ig,23} \\ R_{ig,13} & R_{ig,23} & 1 \end{pmatrix}.$$

All subjects in a given group have the same correlation structure, i.e., $\mathbf{R}_{1g} = \mathbf{R}_{2g} = \ldots = \mathbf{R}_{n_g g}$ for all $g = 1, \ldots, G$.

The simulation study includes various cases of the exchangeable and unstructured correlation structures. Table 1 shows the correlation structures considered here. The first 4 cases have exchangeable correlation structures and the last four cases have unstructured correlation structures.

Table 1: Correlation structures for model (6)

index	$vec(\mathbf{R}_{ig}) = (R_{ig,12}, R_{ig,13}, R_{ig,23})$
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)

The simulation study takes samples from the joint distribution of $(Y_{ig,1}, Y_{ig,2}, Y_{ig,3})$, where $Y_{ig,j}$ indicates whether subject i in group g selects item j. If a subject selects item j, then $Y_{ig,j} = 1$; otherwise, $Y_{ig,j} = 0$. The marginal distributions of $\{Y_{ig,j}, \forall j\}$ need to satisfy model (6) and the correlations between Y_{ig,j_1} and Y_{ig,j_2} for all $j_1 \neq j_2 = 1, 2, 3$ have to meet the chosen correlation structure. For model (6), the covariates X_{igj} were drawn from a standard normal distribution and were fixed in advance for all simulations.

In order to compute the joint distribution, first we calculate the pairwise probability $\Pr(Y_{ig,j_1} = s, Y_{ig,j_2} = t)$ for $s, t = 0, 1$ from the correlation between Y_{ig,j_1} and Y_{ig,j_2} and the marginal probabilities $\pi_{j_1|ig} = \Pr(Y_{ig,j_1} = 1)$ and $\pi_{j_2|ig} = \Pr(Y_{ig,j_2} = 1)$. Since $\mathrm{Corr}(Y_{ig,j_1}, Y_{ig,j_2}) = \mathrm{Cov}(Y_{ig,j_1}, Y_{ig,j_2})/\ (\mathrm{Var}(Y_{ig,j_1})^{1/2}\ \mathrm{Var}(Y_{ig,j_2})^{1/2})$ and $\mathrm{Cov}(Y_{ig,j_1}, Y_{ig,j_2}) = \Pr(Y_{ig,j_1} = 1, Y_{ig,j_2} = 1) - \pi_{j_1|ig}\pi_{j_2|ig}$, we can obtain the pairwise probability $\Pr(Y_{ig,j_1} = 1, Y_{ig,j_2} = 1)$. Then we can find the other pairwise probabilities $\Pr(Y_{ig,j_1} = s, Y_{ig,j_2} = t)$ for $(s,t) = (1,0), (0,1), (0,0)$ from $\Pr(Y_{ig,j_1} = 1, Y_{ig,j_2} = 1)$, $\pi_{j_1|ig}$ and $\pi_{j_2|ig}$.

Finally, we compute joint distributions ($Y_{ig,1}$, $Y_{ig,2}$, $Y_{ig,3}$) (multinomial distribution with 2^c outcomes for each setting ig) from the complete pairwise distributions for all pairs of items using the iterative proportional fitting algorithm. The generation of the joint distributions subject to the marginal distribution satisfying model (6) is analogous to the one applied in the simulation study by Bilder, Loughin & Nettleton (2000).

We draw n (= 50 and 200) observations $\mathbf{y}_{ig} = (y_{ig,1}, y_{ig,2}, y_{ig,3})^T$ randomly from either of the $G = 4$ groups according to the joint probability distributions ($Y_{ig,1}$, $Y_{ig,2}$, $Y_{ig,3}$). We require $n_g > 5$ to achieve better convergence, because the groupwise method is not applicable for small group sizes n_g. Then we fit model (6) by GEE, using the standard method and the groupwise method with $G = 4$, respectively.

Table 2 shows the simulation results for the GEE method for $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T = (0.1, 0.2, 0.3)^T$. The first column shows the total sample size n. The second column shows the index of the true correlation structure $\mathbf{R}_{ig}$ for each of the g groups, where the index is defined in Table 1. For example, if $\mathbf{R}_{i2} = 4$, then the second group has an exchangeable correlation structure with $\alpha = 0.5$.

We define the relative efficiency $RE(\hat{\boldsymbol{\beta}})$ of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_c)^T$ as

$$RE(\boldsymbol{\beta}) = \frac{\sum_{j=1}^{c} \mathbb{E}(\hat{\beta}_j^{TRUE} - \beta_j)^2}{\sum_{j=1}^{c} \mathbb{E}(\hat{\beta}_j - \beta_j)^2} = \frac{\sum_{j=1}^{c} m.s.e.(\hat{\beta}_j^{TRUE})}{\sum_{j=1}^{c} m.s.e.(\hat{\beta}_j)},$$

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. To obtain $\hat{\beta}_j^{TRUE}$, we use the correct correlation of the simulated distribution instead of the correct "working" correlation. This ensures that $\hat{\beta}_j^{TRUE}$ has the smallest mean square error. The relative efficiency of 1.00 is the highest value. Table (2) shows the relative efficiency $RE(\hat{\boldsymbol{\beta}})$ when the model is fitted using the working correlation structure as unstructured (denoted by "unstr"), exchangeable ("exch") and independence ("ind"), respectively.

For the groupwise method, the correlation parameters vary for different groups. For instance, there are 4 parameters for the exchangeable structure. Unlike the groupwise method, the standard method assumes that the correlation parameters remain the same for different groups. Therefore, there is only 1 parameter for the exchangeable structure.

We simulated $10,000$ data sets for each of the configurations. In comparing the performance, a * symbol in the table indicates that the estimators are most efficient among these methods. The results show that for the exchangeable correlation structure, the groupwise method gives more efficient parameter estimators provided different groups have different correlation parameters (e.g., correlation structure = (1, 1, 4, 4) or (1, 2, 3, 4)). The estimators are more efficient when the sample size is larger. The advantage of choosing the groupwise method using unstructured correlation is not obvious when the true correlation parameters vary for different groups. One reason is that for the sample size of 50, we don't have enough observations to estimate 3 correlation parameters for each group, because on average each group has only about 12 subjects. The relative efficiency increases when the sample size in-

Table 2: Relative efficiency ($RE(\hat{\boldsymbol{\beta}})$) with $\boldsymbol{\beta} = (0.1, 0.2, 0.3)^T$ for multiple response data using correlation structures independence (ind), unstructured (unstr) and exchangeable (exch), with both the standard and groupwise methods ($G = 4$)

	correlation structure	working correlation				
		standard method			groupwise method	
n	$\mathbf{R}_{i1}, \mathbf{R}_{i2}, \mathbf{R}_{i3}, \mathbf{R}_{i4}$	unstr	exch	ind	unstr	exch
50	4, 4, 4, 4	0.958	0.983*	0.662	0.907	0.964
50	1, 1, 4, 4	0.819	0.843	0.759	0.884	0.948*
50	1, 2, 3, 4	0.873	0.898	0.814	0.863	0.937*
50	5, 5, 5, 5	0.959*	0.895	0.776	0.833	0.844
50	5, 5, 6, 6	0.948*	0.868	0.693	0.834	0.820
50	5, 6, 7, 8	0.946*	0.906	0.762	0.870	0.896
200	4, 4, 4, 4	0.993	0.999*	0.708	0.975	0.995
200	1, 1, 4, 4	0.870	0.874	0.784	0.973	0.991*
200	1, 2, 3, 4	0.912	0.918	0.826	0.966	0.989*
200	5, 5, 5, 5	0.992*	0.930	0.823	0.971	0.922
200	5, 5, 6, 6	0.986*	0.915	0.765	0.967	0.912
200	5, 6, 7, 8	0.979*	0.930	0.799	0.969	0.944

*: The most efficient estimators.

creases to 200. The groupwise efficiencies for the unstructured correlation are similar to the ones in the standard method. The independent correlation structure has the lowest efficiency across all different configurations.

Generally, we suggest the groupwise method if the exchangeable working correlation structure is chosen. The more parameters the working correlation requires, the bigger the group sizes have to be to gain efficiency advantages from the groupwise method over the standard method. If the unstructured correlation is used, we suggest the standard method, unless the number of subjects in each group is very large. Similar results hold for the repeated multiple response with $T > 1$ (Suesse 2008).

6 Example

Students of a statistics course at the Victoria University of Wellington in New Zealand were asked by the second author to complete a questionnaire (in Appendix B) on 3 different occasions: March 2004, July 2005 and October 2005. The students were asked to tick their favourite bar and also to provide some personal information, such as work status, smoker/non-smoker, etc. The student's sex, major, and age were given by the student record. Also, the study recorded the features of the bars separately. The aim is to get insight into the relationship between the choice of favourite bar based on its features and possible covariates, such as sex, age, etc. The data have $T = 3$ time points and have various bar features to be considered as items. Here, we only consider $c = 3$ items, namely "drink deals" (item 1), "pool table" (item 2) and "sports TV" (item 3). We assign a positive response at occasion t for item j (e.g. "drink deals"), when the student's favourite bar at occasion t has a feature j (e.g. "drink deals").

For the marginal model, we fit model (1) using the logit link. Based on the backward elimination model selection which removes non-significant terms one by one, the final model includes factors "work" (working=1/ not working=0, Question 13), "friends" (yes=1/ no=0, Question 5a), "sex" (male=1/ female=0), "pool" (yes=1/ no=0, Question 4), and "smoker" (yes=1/ no=0, Question 12). Also, the final model has a common effect $\beta_j = \beta_{1j} = \cdots = \beta_{Tj}$ across different time points, with the form

$$\log\left(\frac{\pi_{j|it}}{1-\pi_{j|it}}\right) = \alpha_j + \mathbf{X}_{i0}\boldsymbol{\beta}_{0j} + \mathbf{X}_{it}\boldsymbol{\beta}_j.$$

In the final model, the factors "work", "friends", and "sex" are significant for item 1; "pool" and "sex" are significant for item 2; and "smoker" and "sex" are significant for item 3.

To demonstrate the generalised linear mixed models in Section 4, we use the same covariates as in the marginal final model, with the form

$$\log\left(\frac{\pi_{j|it}}{1-\pi_{j|it}}\right) = \alpha_j + \mathbf{X}_{i0}\boldsymbol{\beta}_{0j} + \mathbf{X}_{it}\boldsymbol{\beta}_j + u_{ij}, \tag{7}$$

where $\mathbf{u}_i = (u_{i1}, \ldots, u_{ic})^T$. The u_{ij} is a random effect for subject i and item j. We assume that the random effect vector $\mathbf{u}_i \in \mathbb{R}^c$ follows a multivariate normal distribution. In the literature, as in Agresti & Liu (2001), the mixed model often includes one single univariate random effect ($\{u_i\}$) to account for dependency between items and occasions, but this seems too stringent. On the contrary, allowing cT correlated random effects (e.g., $\{u_{ijt}, \forall j = 1, \ldots, c; \forall t = 1, \ldots, T\}$), one for each component, does not seem appropriate either due to the large numbers of variance and covariance parameters in $\mathbf{\Sigma}$. For this example, because we expect that the $\pi_{j|it}$ vary more over items than over time, so we choose the random effects that allow a general covariance structure among different items and that allow the structure to remain the same across different time points.

Table 3 shows the parameter estimates for both GEE and GLMM approaches. For GEE, the table includes various correlation structures, such as "independent", "unstructured", "exchangeable", "unstr(item) – exch(time)", and "exch(item) – unstr(time)" for the standard method. We also include the groupwise method (denoted by G) using $G = 2$ groups formed by variable "sex". The structure "unstr(items) – exch(time)" uses the unstructured setting for the correlation between items and the exchangeable setting for the correlation between time points. Similarly, "exch(items) – unstr(time)" uses the exchangeable setting for the correlation between items and the unstructured setting for the correlation between time points. Unfortunately, GEE groupwise method only converges for "exch(items) – unstr(time)", not for "unstr(items) – exch(time)". Table 3 also shows the GEE groupwise method when the correlation structure is "exchangeable" across all items and time points. For the groupwise method, the variable "pool" has a smaller p-value than the p-value for the standard method. For instance, for the structure "ex(items) – unstr(time)", the standard method yields the p-value 0.168 (not significant) and the groupwise method gives a p-value of 0.052 that reaches a moderate significance level.

In summary, we discuss the parameter estimates for structure "exch(items) – unstr(time)" with $G = 2$. 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, given other factors at a fixed level. 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 $\exp(0.553) = 1.77$ times those for non-working people. Similarly, the odds for people who go out to socialize choosing drink deals becomes $\exp(0.575) = 1.78$ times those who do not. 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 do not. 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. 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.

For the generalised linear mixed model (GLMM), we applied the Monte-Carlo-Newton-Raphson (MCNR) algorithm (McCulloch 1997) in combination with confidence regions (Booth & Hobert 1999). We also fitted the model with penalised quasi likelihood (PQL). For the GLMMs, the PQL methods can perform poorly relative to maximum likelihood (McCulloch 1997). Table 3 gives quite different parameter estimates for PQL compared to the MCNR algorithm, indicating a possible bias and yielding unreliable estimates. Since the GLMM is a subject-specific model, we describe the parameter estimates at the subject level. For instance, the odds of choosing a bar that offers drink deals for a person who is working are $\exp(0.544) = 1.72$ (using MCNR) times those for the person who is not working. It compares the odds of choosing a bar that offers drink deals for each person when the person's working status changes. In general, a random effects model with logit link does not imply a marginal model with the logit link. The marginal model parameter estimates are based on the population–averaged level. For the simple case with single univariate random effect, there are approximate relationships between their parameters (Zeger et al. 1988). Agresti & Liu (2001) gave the discussion between the marginal and random effects models when $T = 1$. The same arguments also hold for the case with $T > 1$. Therefore, the parameter estimates are different from those in the GEE method.

7 Discussion

This article mainly focuses on GEE and GLMM methods for modelling repeated multiple responses due to the impractical nature of the ML approach. The ML estimation does not need any assumption about correlation parameters, however, the method becomes infeasible even for small c and T, because data are often highly sparse according to the 2^{cT} possible profiles. The mixed models take into account the dependence among items and time points through the distribution of random effects. It has relatively few parameters compared to the ML method which assumes the multinomial distribution for the 2^{cT} possible profiles. However, the mixed model as in model (7) implies the nonnegative associations across different time points due to the simple structure of the joint distribution. This might not be the case, that is, subjects who respond positively to one item at one time point may not be likely to respond positively to the item at another time point. Furthermore, the mixed model that contains one single univariate random effect (e.g., u_i), implies the nonnegative association across all items and time points.

The marginal models using GEE approach do not assume any subject–specific joint distributions. They use only a working correlation structure for the responses across items and time points to improve efficiency. In general, the GEE method is widely implemented in all common statistical packages. If one wishes to obtain even more efficient estimates, we recommend implementing the GEE procedure using the groupwise method that allows different correlation parameters for different groups. Based on the simulation study, the estimators are more efficient when the correlation structure is exchangeable. Similar results hold for the repeated multiple response case with $T > 1$ (Suesse 2008). One can also consider more sophisticated correlation, such as the autoregressive structure across different time points and the exchangeable structure across different items when $T > 1$.

Although both GEE and GLMM methods seem similar and contain the same fixed effect parameters $\boldsymbol{\beta}$, one does not imply the other. For our example, we are interested in how the probability of choosing a bar having a certain feature depends on different factors. Therefore, the overall (population–averaged) rates are more relevant. Generally speaking, the marginal models seem to be more useful in many applications than the subject–specific models. The subject–specific models might be useful in medical studies, when the effects of interest are at the subject–level. For instance, does the probability of recovery depend on the treatments

Table 3: Parameter estimates, (s.e.), p-value for GEE and GLMM models

method	Drink Deals work	Drink Deals friends	Drink Deals sex	Pool Table pool	Pool Table sex	Sports TV smoke	Sports TV sex
GEE	0.488	0.013	-0.641	-0.843	-0.473	0.201	-0.609
unstr(items)	(0.286)	(0.512)	(0.329)	(0.498)	(0.355)	(0.212)	(0.385)
-exch(time)	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
exch(items)	(0.241)	(0.295)	(0.323)	(0.408)	(0.329)	(0.188)	(0.393)
-unstr(time)	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
exch(items)	(0.215)	(0.323)	(0.324)	(0.377)	(0.335)	(0.196)	(0.401)
-unstr(time)	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
exch ($G = 2$)	(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
exch	(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

and other covariates conditional on patients?

The proposed groupwise method is a form of modelling of the correlation parameters. A common way to model a measure that describes the dependence between items, e.g. odds ratio or correlation parameter, is to express the model as a set of GEE. In this way, we obtain a second set of GEE, in addition to the first set of GEE that describes the mean response model. When this second set is orthogonal to the first set, the fitting is called GEE1 (Prentice 1988) and if both sets are fit jointly then it is called GEE2 (Zhao & Prentice 1990). The proposed groupwise method is a simple method which only need one set of GEE, the set that specifies the mean response model, and therefore provides an alternative to GEE1 and GEE2. For further details of GEE1 and GEE2, see Prentice & Zhao (1991),Lipsitz, Laird & Harrington (1991) and Liang, Zeger & Qaqish (1992).

Finally, we discuss the issue about missing data, which occur in our example. The GEE method assumes data being missing completely at random (MCAR). However, under the weaker assumption of missing at random (MAR), GEE does not provide consistency in contrast to ML methods provided by Lang & Agresti (1994) and Lang (1996). On the other hand, the procedure in GLMMs assumes MAR. For our example, the GEE method seems reasonable, because a subcase of MCAR allows missingness to depend on the observed covariates, e.g. time, major, or sex. It is called the covariate-dependent missingness (Hedeker & Gibbons 2006). However, for other examples, such covariate dependence can be ruled out. Fitzmaurice, Molenberghs & Lipsitz (1995) and Ali & Talukder (2005) considered missing data mechanisms for longitudinal binary data deriving weighted generalised estimation equations (WGEE), an extension of GEE which can handle MAR. Future research will extend our method to the MAR case.

A Proof of Theorem 3.1

Liang & Zeger (1986) showed the following theorem for the property of the GEE estimators:

Theorem A.1 ("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

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

The only difference between Theorems A.1 and 3.1 is condition (1.) and $\mathbf{V}_i$. In Theorem A.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 3.1 matrix $\mathbf{R}_i$ stands for different cluster lengths but also stands for different correlations depending on which group g observation i belongs to.

Now we want to prove Theorem 3.1, 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 in the proof for Theorem A.1 on page 2 in Liang & Zeger (1986). But now we can re-write B^* , given on page 21 in Liang & Zeger (1986), 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 3.1 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 on page 21 in Liang & Zeger (1986).

B Questionnaire

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. Each bar's features were collected separately.

References

Agresti, A. (2002), *Categorical Data Analysis*, Wiley Series in Probability and Statistics, 2nd edition edn, Wiley.

Agresti, A. & Liu, I. (2001), 'Strategies for modelling a categorical variable allowing multiple category choices', *Sociological Methods & Research* **29**(4), 403–434.

Agresti, A. & Liu, I. M. (1999), 'Modelling a categorical variable allowing arbitrarily many category choices', *Biometrics* **55**(3), 936–943.

Ali, M. W. & Talukder, E. (2005), 'Analysis of longitudinal binary data with missing data due to dropouts', *Journal of Biopharmaceutical Statistics* **15**(6), 993–1007.

Bilder, C. R. & Loughin, T. M. (2002), 'Testing for conditional multiple marginal independence', *Biometrics* **58**(1), 200–208.

Bilder, C. R. & 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. & Nettleton, D. (2000), 'Multiple marginal independence testing for pick any/c variables', *Communications in Statistics-Simulation and Computation* **29**(4), 1285–1316.

Booth, J. G. & 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.

Breslow, N. E. & Clayton, D. G. (1993), ‘Approximate inference in generalized linear mixed models’, *Journal of the American Statistical Association* **88**(421), 9–25.

Breslow, N. E. & Lin, X. H. (1995), ‘Bias correction in generalized linear mixed models with a single component of dispersion’, *Biometrika* **82**(1), 81–91.

Fahrmeir, L. & Tutz, G. (2001), *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer series in statistics, 2nd edn, Springer, New York.

Fitzmaurice, G. M., Molenberghs, G. & 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.

Goldstein, H. (1991), ‘Nonlinear multilevel models, with an application to discrete response data’, *Biometrika* **78**(1), 45–51.

Hartzel, J., Agresti, A. & Caffo, B. (2001), ‘Multinomial logit random effects models’, *Statistical Modelling* (1), 81–102.

Hedeker, D. & Gibbons, R. (2006), *Longitudinal Data Analysis*, J. Wiley, Hoboken, NJ.

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. (2005), ‘Homogeneous linear predictor models for contingency tables’, *Journal of the American Statistical Association* **100**(469), 121–134.

Lang, J. B. & Agresti, A. (1994), ‘Simultaneously modelling joint and marginal distributions of multivariate categorical responses’, *Journal of the American Statistical Association* **89**(426), 625–632.

Liang, K. Y. & Zeger, S. L. (1986), ‘Longitudinal data-analysis using generalized linear models’, *Biometrika* **73**(1), 13–22.

Liang, K. Y., Zeger, S. L. & Qaqish, B. (1992), ‘Multivariate regression-analyses for categorical data’, *Journal of the Royal Statistical Society Series B-Methodological* **54**(1), 3–40.

Lipsitz, S. R., Laird, N. M. & 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.

Liu, I. & Suesse, T. (2008), ‘The analysis of stratified multiple responses’, *Biometrical Journal* **50**(1), 135–149.

Loughin, T. M. & Scherer, P. (1998), ‘Testing for association in contingency tables with multiple column responses’, *Biometrics* **54**, 630–637.

McCullagh, P. & 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.

Preisser, J. S. & 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. & Zhao, L. P. (1991), 'Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses', *Biometrics* **47**(3), 825–839.

Raudenbush, S. W., Yang, M. L. & 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.

Schall, R. (1991), 'Estimation in generalized linear-models with random effects', *Biometrika* **78**(4), 719–727.

Stiratelli, R., Laird, N. & Ware, J. H. (1984), 'Random-effects models for serial observations with binary response', *Biometrics* **40**(4), 961–971.

Suesse, T. (2008), Analysis and Diagnostics of Categorical Variables with Multiple Outcomes, PhD thesis, Victoria University of Wellington.

Wedderburn, R. (1974), 'Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method', *Biometrika* **61**, 439–447.

Zeger, S. L., Liang, K. Y. & Albert, P. S. (1988), 'Models for longitudinal data - a generalized estimating equation approach', *Biometrics* **44**(4), 1049–1060.

Zhao, L. P. & Prentice, R. L. (1990), 'Correlated binary regression using a quadratic exponential model', *Biometrika* **77**(3), 642–648.