University of Wollongong
Research Online

Centre for Statistical & Survey Methodology
Working Paper Series

Faculty of Engineering and Information Sciences

2011

Modelling Strategies for Repeated Multiple Response Data

Thomas F. Suesse
University of Wollongong, tsuesse@uow.edu.au

Ivy Liu
Victoria University of Wellington, NZ

Recommended Citation
Suesse, Thomas F. and Liu, Ivy, Modelling Strategies for Repeated Multiple Response Data, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 04-11, 2011, 37p.
http://ro.uow.edu.au/cssmwp/75

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

04-11

Modelling Strategies for Repeated Multiple Response Data

Thomas Suesse and Ivy Liu

Copyright © 2008 by the Centre for Statistical & Survey Methodology, UOW. Work in progress, no part of this paper may be reproduced without permission from the Centre.

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Modelling Strategies for Repeated Multiple Response Data

Thomas Suesse[1] and Ivy Liu[2]
[1]*School of Mathematics and Applied Statistics*
University of Wollongong, Australia
E-mail: tsuesse@uow.edu.au
[2] *School of Mathematics, Statistics and Operations Research*
Victoria University of Wellington, New Zealand

Summary

Agresti and 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 consider a generalised estimating equations (GEE) method to account for different correlations over time and between items as an alternative to standard GEE, which only allow relatively simple correlation structures. We illustrate the different approaches using The Household, Income and Labour Dynamics in Australia (HILDA) Survey, a household-based panel study.

Key words: Multiple Responses, Repeated Measurements, Generalised Linear Models (GLM), Generalised Estimating Equations (GEE), Generalised Linear Mixed Models (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, we use The Household, Income and Labour Dynamics in Australia (HILDA) Survey. It is an annual survey beginning in 2001, that collects information about economic and subjective well-being, labour market dynamics and family dynamics, asking respondents about their daily, weekly and annual expenses. For annual expenses they are asked to "tick all that apply" of the following categories: a) holidays and holiday travel costs, b) private health insurance, c) other insurance (such as home and contents and motor vehicles), d) fees paid to doctors, dentists, opticians, physiotherapists, chiropractors and any other health practitioner, e) Medicines, prescriptions and pharmaceuticals (include alternative medicines), f) Electricity bills, gas bills and other heating fuel (such as firewood and heating oil), g) repairs, renovations and maintenance to your home, h) Motor vehicle repairs and maintenance (include regular servicing), etc. Each outcome category is referred to as an *item* (Agresti & Liu, 1999).

Various authors have considered the analysis of multiple responses. 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, generalized estimating

equations (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 surveyed on several occasions.

Agresti & Liu (2001) treated the responses for each of the items as binary responses (being selected or not). They modelled these correlated responses using the marginal model approach and the mixed model approach. This paper discusses methodologies when multiple responses are recorded on repeated occasions, by treating the responses for each of the items as binary responses. However, unlike Agresti & Liu (2001), our binary responses are correlated in two levels, across both items and different time points. These methods can be generalised to more than two levels, as for HILDA where responses are correlated within households as well.

Section 2 considers the marginal model approach for repeated multiple response data. We primarily focus on GEE (Liang & Zeger, 1986) and consider a variety of possible correlation structures. Standard GEE methods only allow a few options for the correlation structure; these options are unlikely to present a good correlation model for repeated multiple response data. Due to the dependency across several levels, we propose an alternative method combining the levels in a way which allows the use of standard GEE methods while also accounting for multiple correlated levels. A simulation study confirms efficiency advantages of the proposed method.

In Section 2, we also review some of the goodness-of-fit (GOF) statistics and model diagnostics to check the marginal model fit by GEE. Standard likelihood-based methods, such as the deviance, cannot be applied to the GEE method because it is based on quasi-likelihood. Section 3 considers a mixed model approach and reviews some popular model fitting techniques. In Section 4, a simulation study is conducted to investigate the performance of the proposed GEE method to account for different correlated levels. Section 5 illustrates the methods on the HILDA survey using waves E, F, G and H (years 2005-2008). The final section finishes with a discussion.

2 Marginal Modelling

2.1 Maximum Likelihood Approach

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 $\mathbf{Y}_i = (\mathbf{Y}'_{i1}, \ldots, \mathbf{Y}'_{iT})'$ denote the ith subject's $2^{c \cdot T}$ response profile for c items and T time points, where $\mathbf{Y}_{it} = (Y_{i1t}, Y_{i2t}, \ldots, Y_{ict})'$. 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 $\boldsymbol{\pi}_i$ and the mean of $\mathbf{Y}_{it}$ as $\boldsymbol{\pi}_{it}$. Let $\mathbf{x}_i = (\mathbf{x}'_{i1}, \ldots, \mathbf{x}'_{iT})'$ be a vector of covariates of the ith subject, where $\mathbf{x}_{it} = (\mathbf{x}'_{i1t}, \mathbf{x}'_{i2t}, \ldots, \mathbf{x}'_{ict})', t = 1, \ldots, T$ is a column vector of covariates.

Consider item j and time point t only. The n binary responses $\{Y_{ijt}, i = 1, \ldots, n\}$ are independent and standard modelling strategies can be applied to model $\pi_{j|it} := \Pr(Y_{ijt} = 1)$, such as logistic regression:

$$\text{logit}(\pi_{j|it}) = \mathbf{x}'_{ijt}\boldsymbol{\beta}_{jt}. \tag{1}$$

Alternatively any other popular link $h(\cdot)$ can be considered, such as the probit link.

To model c items simultaneously for a single time point, Agresti & Liu (2001) considered several modelling strategies. They introduced the marginal model approach that takes the dependence between items on the same subject into account, in a similar fashion to modelling repeated binary responses. Considering c items simultaneously for model (1) with a general link function $h(\cdot)$, a more compact form can be written as

$$h(\boldsymbol{\pi}_{it}) = \mathbf{X}_{it}\boldsymbol{\beta}_t, \tag{2}$$

with $\boldsymbol{\beta}_t = (\boldsymbol{\beta}_{1t}, \ldots, \boldsymbol{\beta}_{ct})'$ and $\mathbf{X}_{it} := \text{Diag}(\mathbf{x}'_{i1t}, \ldots, \mathbf{x}'_{ict})$, where $h(\boldsymbol{\pi}_{it}) = (h(\pi_{1|it}), h(\pi_{2|it}), \ldots, h(\pi_{c|it}))'$. Furthermore, taking different time points into account, the joint model for repeated

multiple responses can be expressed as

$$h(\boldsymbol{\pi}_i) = \mathbf{X}_i\boldsymbol{\beta}, \tag{3}$$

with $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{iT})$, where $h(\boldsymbol{\pi}_i) = (h(\boldsymbol{\pi}_{i1})', h(\boldsymbol{\pi}_{i2})', \ldots, h(\boldsymbol{\pi}_{iT})')'$.

Repeated multiple responses are more complex, because responses $\{Y_{ijt}\}$ are not only corrected over items (j), but also over the time points (t). One would assume the magnitude of correlation decreases as the time between responses increases. There are several fitting techniques for the joint model. Naively, one can assume independence between all items and occasions and then use ordinary software for generalised linear models (McCullagh & Nelder, 1989). However, this does not give proper standard error estimates for the parameter estimators as independence is not a valid assumption.

Alternatively, model (3) 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 a logistic or log link. 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 $h(\cdot)$. The ML method treats the 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 mean model. As 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 (and will often be zero). This sparseness causes problems with the ML fitting algorithm, and is even worse when some covariates are continuous. 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.

This ML approach does not assume a model for any of the higher moments of the underlying multinomial distribution for each i. One could also additionally model the second

order moments through the odds ratio

$$\theta = \frac{\Pr(Y_{ijt}=1, Y_{ij't'}=1)\Pr(Y_{ijt}=0, Y_{ij't'}=0)}{\Pr(Y_{ijt}=1, Y_{ij't'}=0)\Pr(Y_{ijt}=0, Y_{ij't'}=1)}, j \neq j' \text{ and } t \neq t'$$

or the correlation

$$\rho = \Pr(Y_{ijt}=1, Y_{ij't'}=1) - \Pr(Y_{ijt}=1)\Pr(Y_{ij't'}=1).$$

Both approaches lead to a more complicated ML.

Fitzmaurice & Laird (1993) proposed ML estimation of the mean model based on a quadratic exponential model for the joint distribution. The authors applied the iterative proportional fitting algorithm in each step to obtain the joint distribution from the given model parameters. This procedure is even more complex and not applicable for large c and/or T. Note that the estimating equations for the mean model of this likelihood approach are identical to the generalised estimating equations (GEE), considered below.

The complexity of the ML estimation can be reduced by using the dependence ratios as a measure of association (Ekholm et al., 1995, 2000, 2002, 2003). The dependence ratio for $q \geq 2$ binary variables $Y_{i_1}, \ldots, Y_{i_q}$ is defined as

$$\tau_{i_1,\ldots,i_q} = \frac{\Pr(Y_{i_1}=1, \ldots, Y_{i_q}=1)}{\Pr(Y_{i_1}=1)\cdot\cdots\cdot\Pr(Y_{i_q}=1)}.$$

Modelling the mean and all dependence ratios fully describes the joint distribution. Let $\mathbf{y}_i$ be a value for $\mathbf{Y}_i$. A nice feature of the dependence ratio is that $\Pr(\mathbf{Y}_i = \mathbf{y}_i) = \mathbf{W}\boldsymbol{\Psi}_i$, where $\mathbf{W}$ is a matrix containing elements $-1, +1, 0$ and the elements of the vector $\boldsymbol{\Psi}_i$ are simple functions of the q univariate means $\Pr(Y_{i_j}=1)$ and the $2^q - q - 1$ dependence ratios for subject i. Therefore the log-likelihood

$$\sum_{i=1}^{n} \Pr(\mathbf{Y}_i = \mathbf{y}_i) = \sum_{i=1}^{n} \mathbf{W}\boldsymbol{\Psi}_i$$

can be relatively easily computed compared to Lang's method, which needs iterative procedures to compute $\Pr(\mathbf{Y}_i = \mathbf{y}_i)$. Consider the repeated multiple response case, for each i there are $2^{c \cdot T} - 1$ parameters ($c \cdot T$ means and $2^{c \cdot T} - c \cdot T - 1$ dependence ratios) that describe the joint distribution for the ith cluster. Jokinen (2006) proposed some model simplifications for an exchangeable structure for the dependence ratios, such as a first order Markov chain assumption and a Dirichlet distribution assumption. Under the latter, $\Pr(\mathbf{Y}_i = \mathbf{y}_i)$ can be easily computed from the means and bivariate probabilities. Consequently only second order dependence ratios are needed. It makes ML estimation possible for say medium $c \cdot T$. The `R`-package (R-Development-Core-Team, 2006) `drm` can be used to fit such models.

Such an approach is similar to GEE, which uses a working correlation as an approximation to the true correlation of the joint distribution. Therefore we would not regard such an approach as a real ML approach, but as a pseudo-ML method, because it unlikely to describe the joint distribution accurately. The dependence ratio approach was also criticized by Molenberghs & Verbeke (2004), who strongly advocate for the odds ratio as a measure of association, due to symmetry and ease of interpretation.

To summarise, despite its theoretical appeal, the ML approach is not feasible for moderate or large $c \cdot T$.

2.2 Generalized Estimating Equations Approach

Besides the ML approach, Agresti & Liu (1999) proposed another popular fitting procedure using the GEE method (Liang & Zeger, 1986). The GEE method fits the c marginal models, such as model (1), simultaneously and incorporates a chosen correlation structure/model, known as the *working correlation*.

It is an extension of the quasi-likelihood method (Wedderburn, 1974) for multivariate data. Denote $\mathrm{Var}(\mathbf{Y}_i) = \mathbf{f}_i \cdot \phi^{-1}$ 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 of ones of length cT], and the scale or dispersion parameter by ϕ. Suppose the mean model (3) is true, then the GEE estimates are obtained by computing

the root of the generalised estimation equations

$$\mathbf{U} = \sum_{i=1}^{n} \mathbf{M}_i' \mathbf{V}_i^{-1} \mathbf{r}_i(\boldsymbol{\beta}) = \mathbf{0},$$

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 $\mathbf{r}_i(\boldsymbol{\beta}) = \mathbf{y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta})$. 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 $\widehat{\boldsymbol{\beta}}$. One can adjust the standard errors for $\boldsymbol{\beta}$ based on the 'naive' covariance estimator $\boldsymbol{\Omega}_{naive} := (\sum_{i=1}^{n} \mathbf{M}_i' \mathbf{V}_i^{-1} \mathbf{M}_i)^{-1}$ to reflect what actually occurs for the sample data by using the 'sandwich' covariance estimator $\boldsymbol{\Omega}_{robust} := \boldsymbol{\Omega}_{naive} (\sum_{i=1}^{n} \mathbf{M}_i' \mathbf{V}_i^{-1} \mathbf{r}_i \mathbf{r}_i' \mathbf{V}_i^{-1} \mathbf{M}_i) \boldsymbol{\Omega}_{naive}$, also known as the robust variance. If the working correlation is the true correlation, then the naive variance is consistent and equals the robust variance; however if this does not hold, then only the robust variance is consistent, provided the specification of clusters is correct.

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/model is essential to obtain good variance estimates and more efficient parameter estimates for $\hat{\boldsymbol{\beta}}$. For example, the study in Liang & Zeger (1986). Alternatively, when the focus is primarily on efficiency, Pan & Connett (2002) considered several methods for choosing a working correlation subject to minimising the predictive mean squared error (PMSE). When none of the standard structures are true, these methods often choose the independence structure and provide better efficiency than one of the standard structures.

First consider $T = 1$ and $c > 1$. Then the indices j_1 and j_2 of $R_{j_1 j_2}$ refer to two different items. Common correlation structures between items include

- independence (items): $R_{j_1 j_2} = 0$ for all $j_1 \neq j_2$ (0 parameters)

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

Parameter estimates are usually calculated by the method of moments, but any consistent estimation method can be applied. Next consider $T > 1$ and $c = 1$. Two different indices t_1 and t_2 of $R_{t_1 t_2}$ refer to two different occasions. Options include:

- exchangeable (time): $R_{t_1 t_2} = \alpha$ (1 parameter)
- autoregressive (AR) (time): $R_{t_1 t_2} = \rho^{|t_1 - t_2|}$ (1 parameter)
- unstructured (time): $R_{t_1 t_2} = \alpha_{t_1 t_2}$ ($T(T-1)/2$ parameters)
- m-dependence (time): $R_{t_1 t_2} = \alpha_{|t_1 - t_2|+1}$ for $|t_1 - t_2| > m$, otherwise $R_{t_1 t_2} = 0$ (m parameters)

Typical time dependence structures, such as AR and m-dependence are usually preferred, because observations further apart in time are supposed to be less correlated than those closer in time. Most of these working correlations are provided by common statistical packages. For example `R`-package `geepack` (Yan & Fine, 2004) provides all except m-dependence and `R`-package `gee` provides all of these options.

Repeated multiple responses are characterised by items and time, considered as two levels in a multi-level model. The question remains of how to combine these two levels in an appropriate way. Treating the number of items and the number of time points separately, the correlations between two items within the time point t should match one of the typical correlation structures used for items. Similarly the responses on two different time points within one item should have one of the the typical correlation structures used for longitudinal data. In summary, a reasonable approach is one in which marginally the correlation structure for repeated multiple responses should match the structures for repeated binary observations and also for standard multiple response data.

2.2.1 Combining Levels for GEE

Let the correlation for two responses $Y_{ij_1t_1}$ and $Y_{ij_2t_2}$, referring to two items j_1 and j_2 and two time points t_1 and t_2, be denoted by $R_{j_1j_2,t_1t_2}$. Since the AR structure is multiplicative, i.e. $R_{t,t+k+l} = R_{t,t+k} \times R_{t+k,t+k+l}$, we suggest the same approach for combining the two levels:

$$R_{j_1j_2,t_1t_2} = \begin{cases} R_{j_1j_2}, & \text{for } t_1 = t_2 \\ R_{t_1t_2}, & \text{for } j_1 = j_2 \\ R_{j_1j_2} \times R_{t_1t_2}, & \text{otherwise.} \end{cases} \tag{4}$$

Here $R_{j_1j_2}$ refers to the correlation for two items on one time point and $R_{t_1t_2}$ refers to the correlation for one item on two occasions. This approach meets the requirement of matching marginally the correlation structures of repeated measurements and multiple responses.

Another option uses a simpler approach:

$$R_{j_1j_2,t_1t_2} = \begin{cases} R_{j_1j_2}, & \text{for } t_1 = t_2 \\ R_{t_1t_2}, & \text{for } j_1 = j_2 \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Any other approach is also reasonable as long as marginally the correlation matches the structures of items and time and

$$0 \leq R_{j_1j_2,t_1t_2} < \max(R_{j_1j_2}, R_{t_1t_2}).$$

One would expect that the correlation of responses for two items referring to two different time points is smaller than the correlation of responses for two items (or time points) within the same time point (or item).

Unfortunately, when choosing the AR structure for the time dependence in models (4) or (5), we cannot use standard statistical packages, such as `gee` and `geepack`. The package `geepack` allows the user to specify a design matrix for a linear correlation model, and therefore

only m-dependence would be possible, because AR implies a non-linear correlation model. Specifying the design matrix of a linear correlation model might be too complicated for the inexperienced user or one might insist on an AR structure. In this case, we propose another feasible option using existing software packages discussed next.

First, fit the mean model and choose, for example, the AR structure for the time dependence ignoring item dependence and store the estimates of these correlation parameters. Then fit the same mean model again but use an appropriate structure for the items, for example unstructured, and ignore time dependence. For both cases estimation of these correlation parameters will be consistent. With these correlation estimates compute an appropriate working correlation $\mathbf{R}_i$ for repeated multiple responses using equations (4) or (5). Then refit the mean model again, but with the fixed working correlation structure $\mathbf{R}_i$. The option 'fixed' is standard for most GEE packages. The GEE estimates for the mean model will still be consistent, because GEE only requires consistent estimation of correlation parameters.

To investigate the performance of this method, a simulation study was conducted in Section 4 to compare the method with standard working correlation options and the option for which all parameters are estimated jointly.

2.2.2 Group-wise Correlation Estimation for GEE

Suesse (2008) proposed a simple group-wise method, that assumes that responses Y_{ijt} and $Y_{ij't'}$ for subjects i of the same group are equally correlated, but correlations differs for different groups. Grouping could naturally occur through variables such as sex. Usually the GEE method assumes an equal correlation structure for all subjects i for any two responses Y_{ijt} and $Y_{ij't}$. This is a rather unrealistic and crude way to approximate the true correlation structure. Modelling the correlation or alternatively any second order moments has been proposed by many authors; see Zhao & Prentice (1990); Liang et al. (1992); Yan & Fine (2004). This group-wise method is a special case of these more general approaches. A simulation study showed that when the correlation is indeed different for different groups, the group-wise

method yields more efficient mean model estimates compared to the standard method, which assumes equal correlations for all subjects (Suesse, 2008). When all subjects have equal correlations, then the group-wise method is almost as good as the standard method. However the group-wise method only works well when the number of groups is small, the number of subjects per group is reasonably large,for example ≥ 50 and for correlation structures that are characterised by a small number of parameters.

Unfortunately, the group-wise method is not implemented in standard packages. It can be fitted using the package `geepack` by specifying a design matrix for the linear correlation model, although it is more complicated. Here we only want to make the reader aware that modelling the correlations depending on some covariates might better reflect the nature of the data and might be more important than choosing a proper working correlation, which assumes a basic intercept model for each correlation.

2.2.3 Model Diagnostics for GEE

The major disadvantage of the GEE approach is that it is not a ML approach and standard likelihood based model checking diagnostics for GLMs, such as the deviance, cannot be applied. Next we review some existing model checking methods for GEE.

Horton et al. (1999) and Barnhart & Williamson (1998) proposed goodness-of-fit (GOF) tests that can be regarded as extensions of the famous Hosmer & Lemeshow (1980) statistic, which is based on the idea of forming G groups by partitioning the space of covariates. For each group g a parameter γ_g is added to the model

$$h(\boldsymbol{\pi}_i) = \mathbf{X}_i\boldsymbol{\beta} + \sum_{g=1}^{G} I(i \text{ in group } g)\gamma_g. \tag{6}$$

where $I(event)$ is the indicator function, which is one if the *event* is true and zero otherwise. The model (3) is accepted if the null hypothesis

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_G = 0$$

is not rejected. Horton et al. (1999) used a score test which is asymptotically χ^2_{G-1} distributed to test H_0. Barnhart & Williamson (1998) proposed another score test, that also partitions the covariate space into G regions, but adds parameters for the time and interaction effects time×region. Then the score test is applied to test whether all such added parameters are zero. These proposed GOF statistics are asymptotically χ^2 distributed. The degrees of freedom ($d.f.$) do not depend on p in a simple way, such as $n-p$ or similar. A disadvantage of these statistics was illustrated by Hosmer et al. (1997), who showed that six statistical packages gave six different p-values for a well known example, simply because the p-value depends on the partitioning which is handled differently in these six packages. Lee & Qaqish (2004) proposed another GOF statistic for grouped observations where the correlations are estimated within each group.

Pan (2002) derived the asymptotic approximate distribution of the Pearson chi-squared statistic

$$X^2 = \sum_{i,j,t} \frac{(y_{ijt} - \hat{\pi}_{ijt})^2}{\hat{\pi}_{ijt}(1 - \hat{\pi}_{ijr})} \tag{7}$$

when the data are fit by GEE. The distribution of X^2 is not chi-squared due to correlated observations within a cluster but approximately normal with mean $n \cdot c \cdot T$ and complex variance. The result is based on a first order Taylor series approximation (Pan, 2002).

Pan (2001a,b); Pan & Le (2001) and Pan & Connett (2002) considered model selection in GEE, either for the mean model or correlation model. Pan (2001a) suggested a quasi-likelihood under the independence model criterion (QIC) for GEE, based on the quasi-log-likelihood $Q(\boldsymbol{\beta})$ under the independence assumption. Then $Q(\cdot)$ is evaluated at $\hat{\boldsymbol{\beta}}(\mathbf{R})$ that is obtained under the working correlation $\mathbf{R}$. The QIC is defined as

$$\text{QIC}(\mathbf{R}) = -2Q(\hat{\boldsymbol{\beta}}(\mathbf{R})) + 2\text{trace}(\mathbf{P}), \tag{8}$$

with $\mathbf{P} = \boldsymbol{\Omega}^{-1}_{naive}\boldsymbol{\Omega}_{robust}$. The quasi-log-likelihood Q under the independence model is equal to the log-likelihood L for independent observations. Under the independence model, the

standard AIC $= -2L + 2p$ is an approximation for QIC. However as Pan (2001a) noted, this approximation can only be used to check the mean model, not the working correlation, which can only be done using the QIC. Wang & Hin (2009) and Hin & Wang (2009) developed similar approaches based on QIC.

We believe this approach using QIC is only useful for checking the mean model and using the independence correlation model, but it is not useful for any other working correlation. If the working correlation is not the independence model, there is a non-zero term missing in the QIC. Therefore, it is not suitable for any non-independence model.

The term $Q(\hat{\boldsymbol{\beta}}(\mathbf{R}))$ also punishes for deviations from the independence model due to its definition. In (8), the term $Q(\hat{\boldsymbol{\beta}}(\mathbf{R}))$ is of order $N = n \cdot c \cdot T$ (sample size), as is $\mathbf{U}$, which is part of the missing term. The term trace($\mathbf{P}$) in (8) that punishes the independence model if it is not true, is of order p. Therefore QIC seems especially problematic if N is relatively large compared to p, as for the HILDA data set.

Alternatively one might opt for the Rotnitzky & Jewell (1990) criterion:

$$\mathrm{RJ} = \sqrt{(1 - \mathrm{trace}(\mathbf{P})/p) + (1 - \mathrm{trace}(\mathbf{P}^2)/p)}.$$

The working correlation with the smallest RJ should be chosen. Hin et al. (2007) note that neither QIC nor the RJ criterion preformed well in their simulation study using $n = 100$ and cluster size 5 ($N = 500$). However for larger data sets ($N >> 500$) we expect the RJ criterion to perform better due to the limitations of QIC mentioned above.

Other methods include that of Pan & Connett (2002) who select the working correlation based on minimizing the predictive mean squared error (PMSE). The PMSE is evaluated using the bootstrap method (Efron & Tibshirani, 1993). Liu et al. (2009) considered a more sophisticated model diagnostic approach to check the functional form and link function of a covariate for the proportional odds model based on a cumulative residual process. This method was actually derived more generally for GEE and performs much better than the Hosmer-Lemeshow statistic but is very time consuming. Pan et al. (2001) proposed a marginal

model plot to assess the model adequacy in GEE.

Among all the model diagnostic methods mentioned above, a relatively convenient way to check the GEE model is to first fit the model by ordinary GLM routines and use the AIC to select the covariates of the mean model (or alternatively by QIC using the independence model). The advantage of this approach is that standard model selection methods for GLM can be used and the AIC is an approximation to the QIC. Then in a second step the working correlation should be chosen, for which one could follow Pan & Connett (2002) or use the RJ criterion. In a third step the Pearson statistic can be applied to test the overall GOF of the model. If the Pearson statistic is not feasible due to matrix inversion, as for the large data set HILDA, one can use any of the Hosmer-Lemeshow-type GOF statistics.

3 Generalised Linear Mixed Models

The marginal model (3) is called a population–averaged model, which focuses on the marginal distribution of the responses. Instead of assuming a particular joint 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 a subject-specific model, since parameters are defined on the subject level.

Let $\mathbf{u}_i$ be the random effect vector for subject i and let $\mathbf{Z}_i$ be the design matrix for the random effects. 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 assumed independent within subject i, 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

$$h(\boldsymbol{\pi}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i, \tag{9}$$

where $\mathbf{X}_i$, $h(\cdot)$ and $\boldsymbol{\beta}$ have the same form as in model (3). The design matrix $\mathbf{Z}_i$ for the random effects $\mathbf{u}_i$ consists of rows $\mathbf{z}'_{ijt}$ referring to subject i, item j and time point t. 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

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

We can also write

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

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 $l(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y})$

$$l(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y}) = f(\mathbf{Y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\mathbf{Y}|\mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma}) d\mathbf{u} \tag{10}$$

to obtain ML parameter estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. This likelihood function is often called the *marginal likelihood* after integrating out the random effects (Agresti, 2002).

The integral usually cannot be solved analytically and numerical methods must be applied. Gauss-Hermite quadrature methods directly approximate the integral (10). 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.

Several methods for approximating the marginal likelihood are available; see Stiratelli et al. (1984); Schall (1991); Breslow & Clayton (1993); Zeger et al. (1988) and Goldstein (1991). However, most of them can yield poor estimates, in particular for first order expansions (Breslow & Lin, 1995). Raudenbush et al. (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 non-parametric ML (Hartzel et al., 2001). Another popular method is the EM (expectation-maximisation) algorithm implemented here by treating the random effects as unobserved data. Algorithms have been provided by McCulloch (1997) and Booth & Hobert (1999) among others.

A typical multilevel approach is to consider subjects, time and items as levels. The problem with this approach is that the imposed correlation over time does not resemble a typical time dependence structure, such as the autoregressive structure. Often random intercepts are used in model (9), implying non-negative correlations between the responses. Only if $\mathbf{z}'_{ijt}\mathbf{u}_i$ and $\mathbf{z}'_{ij't'}\mathbf{u}_i$ are monotone in opposite directions is the covariance non-positive (Egozcue et al., 2009). That is, negative correlations cannot be modelled by random intercepts or positive design matrices. In our view, constructing negative correlations by specifying $\mathbf{z}_{ijt}$ seems impractical and therefore GLMM are not useful in modelling data with negative correlations.

In general, if the investigator is interested in the relationship between a response variable and covariates for subjects of the whole population, then the GEE approach is preferable, because the probability of success can be easily calculated for known covariates and by applying the inverse link h^{-1} to (2). If, however, one is more interested in inference for the sample and in predicting future observations, then the GLMM approach is preferable, because for each subject in the sample a subject-specific effect is obtained, $\hat{\mathbf{u}}_i$, from the fitting procedure allowing the prediction of future probabilities by applying h^{-1} to formula (9) for a potential set of covariates $\mathbf{x}_{ijt}$.

4 Simulation Study for GEE Approach Combing Levels

To simplify the situation as much as possible we only consider $c = 2$ and $T = 3$. Including two items is sufficiently large even though there is just one correlation coefficient α, because between any two items one might expect a different correlation. Also, $T = 3$ is large enough to illustrate the AR-structure with parameter ρ. Two correlation structures are used, one for the items (unstructured, $\alpha = 0.2$) and one for time points (AR, $\rho = 0.4$) with the joint correlation structure described in (4). It is unlikely that the true correlation between two responses (Y_{ijt} and $Y_{ij't'}$ with $j \neq j'$ and $t \neq t'$) is zero. Therefore the joint correlation structure described in (5) is not considered for the true model. For both items the same AR parameters $\rho = 0.4$ are used. We consider three models A, B and C. Model A is characterized by

$$\text{logit}(\pi_{j|it}) = \beta_0 + \beta_j X + \beta_3 \cdot t$$

with $\beta_0 = -0.5$, $\beta_1 = -0.5$, $\beta_2 = 0.9$, $\beta_3 = -1.2$ and $X \sim N(0, 1)$, allowing both item and time effects. Model B has the form

$$\text{logit}(\pi_{j|it}) = \beta_{0j} + \beta_j X_j$$

with $\beta_{01} = 0.2$, $\beta_{02} = 0.3$, $\beta_1 = 0.5$, $\beta_2 = 0.7$, X_1 discrete with equal probabilities on $\{0, 1, 2, 3, 4, 5\}$, and $X_2 \sim N(0, 1)$, allowing only item effects. Model C has the form

$$\text{logit}(\pi_{j|it}) = \beta_0 + \beta X,$$

where $\beta_0 = -1$ and $\beta = 3$ with $X \sim N(0, 1)$, with no item or time effects. The number of clusters generated for each of the three models is $n = 30, 100, 500$.

To simulate the data, we need to calculate the joint distribution for each $\mathbf{Y}_i$ for the given marginal means $\pi_{j|it}$ and correlations. From the correlations and $\pi_{j|it}$, the pair-wise probabilities $\Pr(Y_{ijt} = 1, Y_{ij't'} = 1)$ are computed, from which a set of $2^{c \cdot T}$ joint probabilities

can be calculated, as in Lee (1993). There are usually many solutions for such a set, provided a feasible solution exists. The iterative proportional fitting algorithm (IPF) of Gange (1995) is applied to obtain such a solution , this is analogous to the simulation study in Bilder et al. (2000).

We will always fit the correct mean models and only investigate several methods for the working correlations, including the joint correlation structure option 1 [see (4)] and option 2 [see (5)]. The methods considered are:

a) unstructured for whole cluster (unstr)

b) exchangeable for whole cluster (exch)

c) independence for whole cluster (ind)

d) option 1, mean model and working correlation estimated jointly (opt1-j)

e) option 1, ρ and α estimated separately before fitting the mean model (opt1-s)

f) same as d), but option 2 (opt2-j)

g) same as e), but option 2 (opt2-s)

h) only item correlation, ignore time points (item)

i) only time correlation, ignore items (time).

Results of the simulation study for the mean model parameters are shown in Table 1. Models are fitted under both options, so that the effect of omitting a small non-zero correlation can be assessed. The table shows the relative mean squared error (RMSE), the mean squared error (MSE) relative to the GEE method using the correct known (fixed) correlation structure to evaluate the relative efficiency, and the coverage for a 95% confidence interval based on the naive variance and on the robust variance (RMSE, naive, robust). The table shows a single value for each model, even though the models A, B and C refer to models

with several parameters, because showing results for all parameters and models separately does not provide more insight. The MSE and coverage were computed as averages over all mean model parameters to obtain a single value.

Table 1: Simulation Results for models A, B and C under option 1 for $n = 30, 100, 500$ - average RMSE, and average coverage of 95% confidence interval based on naive variance followed by robust variance, average is over all mean model parameters

	Fitting	Mean Models		
n	Method	Model A	Model B	Model C
30	unstr	1.043, 91.1, 89.8$^\bullet$	1.102, 91.0, 90.0°	1.209, 89.7, 91.1°
30	exch	0.995, 89.1, 90.4*	1.062, 91.8, 92.4	1.041, 92.8, 92.4*
30	ind	1.039, 88.7, 89.8*	1.070, 89.2, 92.8	1.041, 83.4, 92.4*
30	opt1-j	0.986, 91.3, 90.5*	1.033, 91.4, 92.3	1.036, 89.9, 92.4*
30	opt1-s	0.985, 91.1, 90.4*	1.034, 91.0, 92.3	1.033, 89.2, 92.4*
30	opt2-j	0.990, 91.2, 90.5°	1.039, 92.1, 92.4	1.019, 90.0, 92.3*
30	opt2-s	0.990, 90.9, 90.4*	1.038, 91.7, 92.4	1.047, 89.1, 92.3*
30	item	1.002, 90.1, 89.3*	1.043, 88.7, 87.8	1.056, 86.7, 87.6*
30	time	1.037, 89.2, 89.9*	1.043, 89.3, 91.0	1.057, 86.1, 90.1*
100	unstr	1.136, 92.0, 92.7°	1.044, 94.0, 93.7*	1.199, 91.8, 93.3°
100	exch	1.024, 91.4, 93.2*	1.062, 91.5, 94.4*	1.036, 93.8, 94.2
100	ind	1.035, 90.3, 92.8	1.067, 88.3, 94.6	1.036, 83.1, 94.2
100	opt1-j	1.011, 92.8, 93.2*	1.024, 91.3, 94.5	1.025, 90.1, 94.2
100	opt1-s	1.017, 92.5, 93.1*	1.025, 90.8, 94.5	1.026, 89.5, 94.2
100	opt2-j	1.017, 92.7, 93.1*	1.032, 91.9, 94.5	1.032, 90.0, 94.2*
100	opt2-s	1.026, 92.4, 93.0*	1.032, 91.4, 94.5	1.034, 89.3, 94.2
100	item	1.032, 91.4, 91.4*	1.035, 88.2, 87.9	1.036, 86.6, 87.7
100	time	1.026, 90.9, 92.4	1.040, 88.5, 91.8	1.041, 85.8, 91.1
500	unstr	1.026, 94.5, 94.4*	1.010, 94.7, 94.6	1.045, 94.2, 94.5*
500	exch	1.013, 93.0, 94.7	1.026, 93.1, 94.8	1.014, 94.5, 94.8
500	ind	1.013, 92.2, 94.7	1.032, 90.9, 94.9	1.014, 88.0, 94.8
500	opt1-j	1.006, 94.0, 94.7	1.010, 92.7, 94.8	1.008, 92.1, 94.9
500	opt1-s	1.007, 93.9, 94.7	1.010, 92.4, 94.9	1.009, 91.8, 94.9
500	opt2-j	1.009, 93.9, 94.7*	1.013, 93.0, 94.9	1.009, 92.2, 94.9*
500	opt2-s	1.009, 93.9, 94.7	1.014, 92.8, 94.9	1.012, 91.9, 94.9
500	item	1.013, 93.2, 93.1	1.016, 90.9, 90.8	1.014, 90.4, 90.8
500	time	1.011, 92.7, 93.9	1.019, 91.1, 93.2	1.017, 89.6, 92.6

Non-convergence rate: 0% (no symbol), $0 - 10\%$ (*), $10 - 50\%$ ($^\circ$), $> 50\%$ ($^\bullet$)

The table does not show the confidence interval (CI) length and also excludes the bias. The bias is negligible and the CI length is monotone in the coverage, because the CI is centered around the same $\hat{\boldsymbol{\beta}}$. Therefore, a method with a smaller coverage has a shorter CI.

One might wonder about the efficiency (< 1) for $n = 30$ and model A. This comes from

the fact that the method with the true correlation structure has higher convergence rate than the methods for which the correlation must be estimated. It results in an unequal set of simulated data sets for which the MSE was computed. This is similar to the problem of non-response leading to biased estimates.

Before making some interpretation of the tables, one has to consider how the methods a) - i) were applied. All methods except methods 'item' and 'time' were applied to the whole cluster of size $c \cdot T$, whereas methods h) and i) were only applied to the clusters that defined items and time points, respectively. Therefore these methods wrongly identify the clusters. The robust variance is not consistent for methods 'item' and 'time' due to cluster mis-specification.

Not surprisingly, the larger the number of clusters becomes, the more accurate the robust variance due to the consistency of the robust variance, except with methods 'time' and 'item'. The naive variance seems rather unreliable. Method 'op1-j' seems generally best, which was to be expected, because it uses the true working correlation. The suggested and relatively easily implementable method 'opt1-s', as an alternative to method 'opt1-j', performs almost as well. The difference in relative efficiency is almost negligible, which is to some extent surprising, because we would expect a higher gain in efficiency if the correlation parameters and the mean model parameters are all estimated jointly.

The method 'unstr' usually performs poorly for small n in terms of both relative efficiency and non-convergence, but improves with n. Methods 'opt2-j' and 'opt2-s', which assume zero correlation between responses of the same person for different items and different time-points, are generally worse than methods 'opt1-j' and 'opt1-s'. The naive variance does not generally perform well, even when the working correlation is the true correlation structure. For large n the robust variance is to be preferred and only for small n is the naive variance preferred, but only for a reasonable working correlation.

Table 1 also gives an indication of the bias of standard errors, because under-coverage indicates that standard errors are under-estimated and over-coverage indicates over-estimation

of standard errors. The results show that the suggested method 'op1-s' performs well and should be used in practice if one wishes to use existing software and does not want to implement method 'opt1-j'.

5 Example: The Household, Income and Labour Dynamics in Australia (HILDA) Survey

The data used in this article come from waves E, F, G and H (years 2005-2008) of the Household, Income and Labour Dynamics in Australia (HILDA) Survey. Details are documented in Wooden et al. (2002). In the first wave (wave A, 2001), 7683 households representing 66% of all in-scope households were interviewed, generating a sample of $15,127$ persons who were 15 years or older and eligible for interviews, of whom $13,969$ were successfully interviewed. Subsequent interviews for later waves were conducted about one year apart. In addition to the data collected through personal interviews, each person completing a personal interview was also given a self-completion questionnaire to be returned on completion by mail or handed back to the interviewer at a subsequent visit to the household.

The HILDA survey contains detailed information on economic and subjective well-being, labour market dynamics and family dynamics. Information relating to individuals' health was collected in both the personal interviews and self-completion questionnaires. In the personal interviews, individuals were asked whether they had a long-term condition, impairment or disability that restricted everyday activities and had lasted or was likely to last for six months or more. Examples are shortness of breath, long term mental health condition and pain.

In the self-completion questionnaire, the Short Form 36 (SF-36) asks questions about the health status. The SF-36 is a measure of general health and wellbeing, and produces scores for eight dimensions of health (Ware et al., 2000), such as mental health, general health, physical functioning and vitality. Scores for all scales range from 0 to 100, with higher scores indicating better health.

In the self-completion questionnaire, respondents were also asked about their daily, weekly, monthly and annual expenses. Two of the items for annual expenses are: i) private health insurance (PHI) and ii) fees paid to doctors, dentists, opticians, physiotherapists, chiropractors and any other health practitioner (FD) (often referred to as 'extras'). In Australia, the government provides a compulsory basic health cover for everyone and purchasing a private health insurance as a top-up cover is optional. Therefore respondents might tick none, one or both of the two items. The first item is available from wave E (2005) and the second from wave F (2006). We have access to waves A to H (2001-2008). Therefore $T = 4$ for the first item and $T = 3$ for item number 2. One of the research question governments and private health insurers might be interested in is how these two items relate to various covariates, such as the health scores, the long term health conditions, etc.

HILDA provides a number of such health variables: i) alcohol drinking status (abstainer, ex-drinker, low risk, risky, high risk), ii) health scores (0-100): mental health, general health, physical functioning and vitality from the SF-36, iii) long term health conditions (indicator for developed at previous wave - 'developed T-1', at current wave 'developed T', shortness of breath, pain, mental health, etc.), smoking status (do not smoke, no longer smoke, smoke weekly but less often than daily, less often than weekly), iv) number of cigarettes a week, v) satisfaction scores (0-100) for life and with partner. The analysis accounted also for sex, age, labor force status, race, dependent person (young adult living with parents), household size (1,2,3,4,5,6+), number of children (0,1,2,3+) and education (higher education – masters or doctorate, grad diploma, grad certificate, Bachelor or honours Advanced diploma, diploma, some education – Cert I,II,III or IV, Cert not defined, Year 12, and no education), major statistical region (Sydney, Balance of New South Wales, Melbourne, Balance of Victoria, Brisbane, Balance of Queensland, Adelaide, Balance of SA, Perth, Balance of Western Australia, Tasmania, Northern Territory and Australian Capital Territory) and remoteness area (Major City, Inner Regional Australia, Outer Regional Australia, Remote Australia, Very Remote Australia).

For this example, we consider another correlated level – household. The notation $\pi_{j|ith}$ is the probability of item j was ticked at time t by subject i who was in household h. Using the GEE method, we first select the mean model

$$\text{logit}(\pi_{j|ith}) = \mathbf{x}'_{ijth}\boldsymbol{\beta}_j^{GEE},$$

using AIC under the standard independence model, since this is an approximation to QIC. Here we assume that the effects for all waves are the same. The final joint model consists of $p = 97$ chosen covariates. Then we assess the working correlation model by computing the RJ criterion and trace($\mathbf{P}$).

GEE with the unstructured working correlation did not converge due to the large data set. We use the working correlation referring to (4), denoted by opt1, and the same option but allow different correlation parameters for each item; this method is denoted by opt1*. Because the responses are correlated among three levels: items, time and household, option (4) is extended by combining multiplicatively three levels, not only items and time. We also fit the mean model with the working correlations: independence (ind), exchangeable (ex), only accounting for time dependence (time), for items dependence (item) and for households dependence (HH).

Table 2 shows the results of the RJ criterion and trace($\mathbf{P}$). Both measures say that opt1* is the best choice followed by opt1.

Table 2: Assessing Working Correlation Models for HILDA

	Working Correlation						
Measure	opt1	opt1*	ind	ex	time	item	HH
RJ	217	201	1300	790	240	1405	667
trace($\mathbf{P}$)	581	558	1191	985	574	1231	920

For opt1*, the AR parameters for the two items are 0.44 for FD and 0.90 for PHI. The correlation, between FD and PHI is -0.26 and the correlation between members of the same household is 0.53. For opt1 the single AR parameter is 0.79. The HILDA data set also

contains area information. We did not account for the dependence between people from the same area in the correlation model. However, in the mean model we add a main effect for each of the major statistical regions and remoteness areas of Australia.

Finally to check the overall model for GEE with opt1, we cannot apply the Pearson statistic as suggested by Pan (2002) due to the large data set. Instead, we compute the Hosmer-Lemeshow statistic with 10 groups (Horton et al., 1999) which gave a p-value of 0.25. The final GEE model was accepted, even though one must keep in mind that such tests usually have a low power.

For fitting mixed models, the `R`-package `lme4` (Bates & Maechler, 2010) was applied which uses a Gauss-Hermite quadrature approximation of the marginal likelihood. We consider the following mixed model

$$\text{logit}(\pi_{j|ith}) = \mathbf{x}'_{ijth}\boldsymbol{\beta}_j^{GLMM} + u_h + u_{j|i} + u_{j|h} + u_{ti} + u_{th},$$

assuming these random intercepts are independent of each other. Instead of just accounting for a single random effect, e.g. u_h, this model accounts for several effects, individual level random effects $u_{j|i}$ (subject-item) and u_{ti} (subject-time), and household level random effects u_h (household - intercept), $u_{j|h}$ (household-item) and u_{th} (household-time). These effects allow us to get more insight into the dependence of items across time-points and household members.

The fitting results for a GLM, GEE (opt1) and GLMM are presented in Tables 3 and 4. To preserve space estimates for major statistical region and remoteness area are not shown. All other variables not shown were excluded by the model selection procedure. The analysis of GEE shows that compared to males, females are more likely to pay fees for doctors and extras than to pay for health insurance. It also happens for the mid-age group (35-74) compared to the baseline age group (18-24). Those with alcohol drinking status low risk, risky or high risk (say drinkers) are more likely to pay fees for doctors and extras than to purchase private health insurance compared to abstainers.

There could be many reasons to explain these results. For example, it could be that drinkers might have less money left over or be higher risk takers than non-drinkers, but they might require more frequent medical services to treat medical conditions associated with their drinking status.

Our primary focus of this paper is not on interpretation of such parameters but on the statistical modeling and its influence on the associated p-values. The tables clearly show that p-values of the GLM approach are smallest. This was expected, because GLM does not account for dependence between observations. Parameter estimates between GEE and GLM are not very different, but standard errors and p-values are. GLMM shows a different picture. Fixed effects estimates are usually larger in magnitude, as are standard errors, but p-values are generally similar to those of GEE, even though for particular parameters differences can be quite large. This can be explained by an approximate relationship between a marginal model and a mixed model (Zeger et al., 1988):

$$\mathbf{x}'_{ijth}\boldsymbol{\beta}_j^{GEE} \approx a(\boldsymbol{\Sigma})\mathbf{x}'_{ijth}\boldsymbol{\beta}_j^{GLMM},$$

where $a(\boldsymbol{\Sigma})$ is a constant depending on the random effects estimates and on $\mathbf{Z}_{ijt}$. The variance estimates of the random effects for u_h, $u_{j|i}$, $u_{j|h}$, u_{ti}, u_{th} are $\sigma_h^2 = 13.81$, $\sigma_{j|i}^2 = 11.85$, $\sigma_{j|h}^2 = 8.25$, $\sigma_{ti}^2 = 2.67$ and $\sigma_{t|h}^2 = 0.377$. This gives $a(\boldsymbol{\Sigma}) \approx 0.26$, implying that fixed effect estimates of GLMM are approximately four times larger than those of GEE.

In our example the item correlation between two items estimated by GEE is -0.26, indicating a negative correlation. However, a simple GLMM with intercepts assumes non-negative correlations (see Section 3). To meet the assumption, we applied a trick to obtain responses that are non-negatively correlated by transforming the 0/1 binary response to a 1/0 response for item 1. That is, positive responses become negative and negative responses become positive. Note this transformation changes the sign of the estimates for item $j = 1$, to make them comparable with the GEE method, the estimates were multiplied by -1. For this example this transformation works, but for a general case with several items it might

not work. For example, assume there are 3 items A, B and C, where items A and B and items B and C might be positively correlated, but items A and C are negatively correlated. In this instance, there is no transformation that makes all correlations positive.

6 Discussion

This article mainly focuses on GEE and GLMM methods for modelling repeated multiple responses, because of the impractical nature of the marginal ML approach. Using Lang's method, the ML estimation does not require any assumption about correlation parameters. However, this method and any other method becomes infeasible even for small c and T, because data are often highly sparse due to the 2^{cT} possible profiles. Mixed models take the dependence among items and time points through the distribution of random effects into account. They have relatively few parameters compared to the ML method, which assumes the multinomial distribution for the 2^{cT} possible profiles. However mixed models, such as in model (9), imply non-negative 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 and transformations might not entirely solve this problem, as indicated in Section 5.

The marginal models using the 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 relative efficiency. In general, the GEE method is widely implemented in all common statistical packages and one might use any of the common simple working correlation structures to obtain more efficient mean model estimates compared to the independence model.

If one wishes to obtain even more efficient estimates, we recommend using the correlation model (4) to account for the two types of correlation, the item correlation and time-points

Table 3: Results for fees paid to doctors and extras

Variable	GLM estimate	GLM (s.e.)	GLM p-value	GEE estimate	GEE (s.e.)	GEE p-value	GLMM estimate	GLMM (s.e.)	GLMM p-value
Intercept	1.504	(0.178)	$< 1e-10$	1.598	(0.225)	$< 1e-10$	5.536	(0.602)	$< 1e-10$
Age - baseline: 18-24									
25-34	0.289	(0.064)	$5.6e-06$	0.168	(0.092)	0.0686	0.695	(0.223)	0.0018
35-49	0.382	(0.057)	$< 1e-10$	0.170	(0.077)	0.0277	0.522	(0.202)	0.0097
50-74	0.761	(0.059)	$< 1e-10$	0.286	(0.074)	$1e-04$	1.126	(0.203)	$3.1e-08$
>74	0.307	(0.084)	$3e-04$	−0.057	(0.114)	0.6154	0.131	(0.297)	0.6576
Race - baseline: Not indigenous									
Indigenous	−0.683	(0.117)	$5.2e-09$	−0.542	(0.131)	$3.6e-05$	−2.231	(0.497)	$7.2e-06$
Female	0.449	(0.037)	$< 1e-10$	0.387	(0.04)	$< 1e-10$	1.274	(0.139)	$< 1e-10$
Education - baseline: higher education									
some education	−0.643	(0.047)	$< 1e-10$	−0.566	(0.061)	$< 1e-10$	−1.685	(0.205)	$< 1e-10$
no education	−1.049	(0.049)	$< 1e-10$	−0.923	(0.063)	$< 1e-10$	−2.653	(0.221)	$< 1e-10$
Baseline: English first language learned									
English not first language	−0.186	(0.046)	$6.3e-05$	−0.057	(0.051)	0.2704	−0.198	(0.143)	0.1641
Gross weekly income	0.001	$(4.2e-05)$	$< 1e-10$	0.001	$(3.9e-05)$	$3.4e-05$	0.001	(0.001)	$9e-04$
Labour Force Status - baseline: employed									
unemployed	−0.505	(0.108)	$3.1e-06$	−0.060	(0.140)	0.6678	−0.724	(0.285)	0.0111
Not in the labour force	−0.340	(0.052)	$< 1e-10$	−0.171	(0.065)	0.0083	−0.889	(0.183)	$1.1e-06$
Satisfaction Scores 0-100									
Satisfaction - Life	−0.045	(0.011)	$6.8e-05$	−0.009	(0.013)	0.4869	0.004	(0.033)	0.8997
Satisfaction - Partner	−0.050	(0.014)	$3e-04$	−0.022	(0.017)	0.1884	−0.087	(0.039)	0.0251
Marital Status - baseline: married									
De facto	−0.672	(0.054)	$< 1e-10$	−0.639	(0.079)	$< 1e-10$	−1.891	(0.227)	$< 1e-10$
Separated	−1.053	(0.094)	$< 1e-10$	−0.775	(0.122)	$2.2e-10$	−2.248	(0.344)	$< 1e-10$
Divorced	−0.956	(0.063)	$< 1e-10$	−0.815	(0.09)	$< 1e-10$	−2.386	(0.277)	$< 1e-10$
Widowed	−0.860	(0.07)	$< 1e-10$	−0.838	(0.1)	$< 1e-10$	−2.403	(0.317)	$< 1e-10$
Never married and not de facto	−1.106	(0.054)	$< 1e-10$	−1.007	(0.075)	$< 1e-10$	−3.279	(0.231)	$< 1e-10$
Alcohol drinking status - baseline: abstainer									
ex-drinker	0.232	(0.081)	0.0042	0.135	(0.103)	0.1903	0.798	(0.255)	0.0018
low risk	0.563	(0.061)	$< 1e-10$	0.350	(0.08)	$1.2e-05$	1.339	(0.21)	$1.8e-10$
risky	0.647	(0.086)	$< 1e-10$	0.359	(0.106)	$7e-04$	1.250	(0.279)	$7.3e-06$
high risk	0.409	(0.11)	$2e-04$	0.150	(0.134)	0.2639	0.560	(0.36)	0.1204
Long Term Health Conditions									
Pain	−0.205	(0.071)	0.0037	−0.120	(0.075)	0.1091	−0.493	(0.199)	0.0132
Shortness of Breath	−0.202	(0.081)	0.0125	0.059	(0.085)	0.4886	−0.132	(0.238)	0.5783
Health Scores 0-100									
Physical functioning	0.006	(0.001)	$4.6e-10$	0.003	(0.001)	0.0421	0.02	(0.003)	$< 1e-10$
Bodily Pain	−0.006	(0.001)	$2.7e-09$	−0.004	(0.001)	$2e-04$	−0.017	(0.003)	$5e-10$
Vitality	−0.006	(0.001)	$7.8e-06$	−0.001	(0.002)	0.5946	−0.013	(0.004)	0.0013
Mental Health	0.010	(0.001)	$< 1e-10$	0.002	(0.002)	0.1776	0.013	(0.004)	0.0023
Smoking Status - baseline: do not smoke									
no longer smoke	−0.119	(0.043)	0.0054	−0.041	(0.055)	0.453	−0.124	(0.159)	0.4361
smoke daily	−0.657	(0.066)	$< 1e-10$	−0.39	(0.085)	$4.6e-06$	−1.416	(0.224)	$2.8e-10$
smoke weekly	−0.406	(0.117)	$5e-04$	−0.209	(0.154)	0.1768	−0.786	(0.33)	0.0171
less often than weekly	−0.341	(0.140)	0.0151	−0.3	(0.125)	0.0163	−1.002	(0.41)	0.0146
Number Cigarettes	−0.001	(0.001)	0.0496	0	(0.001)	0.773	−0.002	(0.001)	0.1543

Table 4: Results for Private Health Insurance

	GLM			GEE			GLMM		
Variable	estimate	(s.e.)	p-value	estimate	(s.e.)	p-value	estimate	(s.e.)	p-value
Intercept	0.081	(0.129)	0.5305	-0.764	(0.108)	$< 1e-10$	-1.955	(0.536)	$3e-04$
Age - baseline: 18-24									
25-34	-0.186	(0.052)	$3e-04$	-0.031	(0.046)	0.5027	-0.840	(0.206)	$4.6e-05$
35-49	-0.463	(0.047)	$< 1e-10$	-0.102	(0.036)	0.0043	-1.066	(0.187)	$1.3e-08$
50-74	-1.101	(0.048)	$< 1e-10$	-0.374	(0.037)	$< 1e-10$	-2.482	(0.194)	$< 1e-10$
>74	-0.772	(0.071)	$< 1e-10$	-0.072	(0.055)	0.1962	-0.503	(0.308)	0.1021
Female	-0.198	(0.027)	$< 1e-10$	-0.077	(0.024)	0.0013	-0.440	(0.124)	$4e-04$
Race - baseline: Not indigenous									
Indigenous	0.667	(0.116)	$9.3e-09$	0.585	(0.12)	$1e-06$	1.618	(0.62)	0.0091
Wave	0.022	(0.012)	0.0626	-0.004	(0.006)	0.5232	0.007	(0.035)	0.8461
Dependent Person	-0.289	(0.099)	0.0036	-0.041	(0.099)	0.6802	-0.53	(0.382)	0.1659
Education - baseline: higher education									
some education	0.774	(0.032)	$< 1e-10$	0.665	(0.037)	$< 1e-10$	2.902	(0.167)	$< 1e-10$
no education	0.976	(0.035)	$< 1e-10$	0.863	(0.043)	$< 1e-10$	3.487	(0.188)	$< 1e-10$
Baseline: English first language learned									
English not first language	0.289	(0.034)	$< 1e-10$	0.123	(0.023)	$1e-07$	0.700	(0.13)	$7.1e-08$
Gross weekly income	-0.001	$(3.0e-05)$	$< 1e-10$	0.000	$(2.1e-05)$	$< 1e-10$	-0.001	(0.00)	$< 1e-10$
Labour Force Status - baseline: employed									
unemployed	0.488	(0.104)	$2.7e-06$	0.010	(0.057)	0.86	0.530	(0.338)	0.1167
Not in the labour force	-0.004	(0.039)	0.9154	0.004	(0.029)	0.9032	-0.026	(0.168)	0.8775
Satisfaction Scores 0-100									
Satisfaction - Partner	0.039	(0.008)	$7.7e-07$	0.015	(0.006)	0.0086	0.064	(0.03)	0.0325
Alcohol drinking status - baseline: abstainer									
ex-drinker	0.044	(0.067)	0.5115	0.023	(0.047)	0.6292	-0.072	(0.261)	0.7825
low risk	-0.39	(0.048)	$< 1e-10$	-0.138	(0.039)	$5e-04$	-1.059	(0.204)	$2.1e-07$
risky	-0.473	(0.066)	$< 1e-10$	-0.167	(0.05)	$8e-04$	-0.766	(0.268)	0.0042
high risk	-0.339	(0.09)	$2e-04$	-0.184	(0.063)	0.0038	-0.972	(0.361)	0.0071
Long Term Health Conditions (LTHC)									
Developed T-1	-0.135	(0.092)	0.1393	-0.042	(0.046)	0.3584	-0.206	(0.288)	0.476
Mental health	0.212	(0.132)	0.1079	0.077	(0.072)	0.2878	0.576	(0.457)	0.2081
Shortness of Breath	0.269	(0.069)	$1e-04$	0.025	(0.039)	0.5164	0.658	(0.259)	0.0111
Marital Status - baseline: married									
De facto	0.674	(0.041)	$< 1e-10$	0.609	(0.047)	$< 1e-10$	2.673	(0.202)	$< 1e-10$
Separated	0.915	(0.076)	$< 1e-10$	0.632	(0.069)	$< 1e-10$	3.152	(0.318)	$< 1e-10$
Divorced	1.034	(0.05)	$< 1e-10$	0.813	(0.057)	$< 1e-10$	3.882	(0.263)	$< 1e-10$
Widowed	0.648	(0.057)	$< 1e-10$	0.548	(0.065)	$< 1e-10$	2.676	(0.308)	$< 1e-10$
Never married and not de facto	0.907	(0.044)	$< 1e-10$	0.952	(0.053)	$< 1e-10$	3.91	(0.226)	$< 1e-10$
Smoking Status - baseline: do not smoke									
no longer smoke	0.261	(0.031)	$< 1e-10$	0.116	(0.028)	$3.1e-05$	0.621	(0.14)	$9.9e-06$
smoke daily	0.989	(0.039)	$< 1e-10$	0.53	(0.04)	$< 1e-10$	3.108	(0.19)	$< 1e-10$
smoke weekly	0.633	(0.093)	$< 1e-10$	0.365	(0.061)	$2.3e-09$	1.807	(0.337)	$8.5e-08$
less often than weekly	0.329	(0.106)	0.0019	0.196	(0.068)	0.0039	0.986	(0.381)	0.0096
Health Scores 0-100									
Physical functioning	-0.003	(0.001)	$3.1e-06$	-0.002	(0)	0.0022	-0.01	(0.003)	$4e-04$
Mental Health	-0.006	(0.001)	$< 1e-10$	-0.001	(0.001)	0.0099	-0.008	(0.003)	0.0282

correlation. In addition, one can use the group-wise method that allows different correlation estimates for groups. If one wishes to use the AR structure for the time-points, then standard GEE packages, such as the R-package `geepack` cannot fit such a correlation model, due to the non-linearity of the AR model. As an alternative, we recommend obtaining an estimate for the time-points correlation and then separately obtaining an estimate of the item correlation. Then, we fit the final model with a fixed working correlation specified by (4) using these time and item correlation estimates. This alternative fitting strategy also works for the group-wise method. The simulation study has shown that this method works almost as well as jointly estimating the correlation and mean model parameters. This method is a trick that enables us to use existing software and avoids writing new code, although it is still an option for the experienced user.

There are some advantages of the group-wise GEE method. Suesse (2008) showed that the efficiency of mean model parameters is improved if the correlation between two responses Y_{ijt} and $Y_{ij't'}$ is not the same across different groups and the number of subjects per group is at least 50. When the underlying correlation model is indeed true, parameter estimates can be quite different between the standard and the group-wise method.

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 paying fees to doctors and extras (FD), and paying private health insurance (PHI) depends on different factors; and comparing the effects on FD and PHI. Therefore, the overall (population–averaged) rates are more relevant. Generally speaking, the marginal models seem to be more useful than the subject–specific models in many applications. The subject–specific models might be useful in medical studies, when the effects of interest are at the subject–level. For example, does the probability of recovery depend on the treatments and other covariates conditional on the patient? Or what is the probability for a future item response of a subject, given the subject–specific effect and hypothetical covariates?

Other model approaches not considered here are marginalized GLMM, transition models

and log-linear models; for a good summary see Diggle et al. (2002). Marginalized GLMM have the advantage of a marginal interpretation, like GEE, but also the advantage that the joint distribution follows a GLMM, allowing likelihood-based inference. This approach is useful if, for example, a multi-level model is applied and a marginal interpretation is sought. Transitional models do not only assume that the linear predictor of Y_{ijt} depends on a set of covariates but also on previous observations, e.g. on $Y_{ij,t-1}$. This approach seems more useful than the GLMM approach when the main goal is prediction of future observations. To apply this approach for repeated multiple response data and to make items dependent, one could assume that the linear predictor of $Y_{i,j_1,t}$ depends also on $Y_{i,j_2,t}$ with $j_1 \neq j_2$. Log-linear models seem least useful for such complex data, because marginalization and fitting becomes increasingly complex for large $c \cdot T$, as with ML estimation for marginal models discussed in Section 2. We did not consider these approaches here in detail, because in contrast to GEE and GLMM, most statistical packages do not offer to fit such models. Each of the approaches is also similar to the GEE or the GLMM approach. Hence the extension of the proposed models to repeated multiple response data is straightforward.

Finally, we discuss the issue about missing data, which occur in our example. The GEE method assumes data being missing completely at random (MCAR). Under the weaker assumption of missing at random (MAR), GEE does not provide consistency in contrast to ML methods provided by Lang & Agresti (1994), Lang (1996) and Lang (2005). On the other hand, the procedure in GLMMs only requires MAR. However, for our example, the GEE method seems reasonable, because a sub-case of MCAR allows missingness to depend on the observed covariates, e.g. time-point, age or sex. It is called the covariate-dependent missingness (Hedeker & Gibbons, 2006). For the general MAR case, Fitzmaurice et al. (1995) and Ali & Talukder (2005) considered a missing data mechanisms for longitudinal binary data deriving weighted generalized estimation equations (WGEE), an extension of GEE.

As illustrated for the HILDA survey in which we accounted for households, items and time-points, the structure of the correlation model (4) and the idea of separate fitting can

be applied to any multi-level-type data accounting for different levels. Future research might focus on empirical studies that investigate how the correlations between responses on the same subject for which items and time-points are different can be modeled. Model (4) suggests using a product, but other functions might be more appropriate. Also, an `R`-package can be provided to allow the joint fitting procedure using the existing GEE packages, as suggested in this article. In the future, the user does not need to implement it 'by-hand'.

Bibliography

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

Agresti, A. & Liu, I. (2001). Strategies for modeling a categorical variable allowing multiple category choices. *Sociol. Methods. Res.*, **29**(4), 403–434.

Agresti, A. & Liu, I. M. (1999). Modeling 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. *J. Biopharm. Stat.*, **15**(6), 993–1007.

Barnhart, H. X. & Williamson, J. M. (1998). Goodness-of-fit tests for gee modeling with binary responses. *Biometrics*, **54**(2), 720–729.

Bates, D. & Maechler, M. (2010). R-package lme4: Linear mixed-effects models using S4 classes.

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. *Commun. Stat.-Simul. Comput.*, **29**(4), 1285–1316.

Booth, J. G. & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, **61**, 265–285.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.*, **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.

Diggle, P., Heagerty, P. J., Liang, K. Y., & Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, 2nd edition.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Egozcue, M., Garcia, L., & Wong, W. (2009). On some covariance inequalities for monotonic and non-monotonic functions. *J. Inequal. Pure Appl. Math.*, **10**(3), 1–16.

Ekholm, A., Jokinen, J., & Kilpi, T. (2002). Combining regression and association modelling for longitudinal data on bacterial carriage. *Stat. Med.*, **21**(5), 773–791.

Ekholm, A., Jokinen, J., McDonald, J. W., & Smith, P. W. F. (2003). Joint regression and association modeling of longitudinal ordinal data. *Biometrics*, **59**(4), 795–803.

Ekholm, A., McDonald, J. W., & Smith, P. W. F. (2000). Association models for a multivariate binary response. *Biometrics*, **56**(3), 712–718.

Ekholm, A., Smith, P. W. F., & McDonald, J. W. (1995). Marginal regression analysis of a multivariate binary response. *Biometrika*, **82**(4), 847–854.

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

Fitzmaurice, G. M. & Laird, N. M. (1993). A likelihood-based method for analyzing longitudinal binary responses. *Biometrika*, **80**(1), 141–151.

Fitzmaurice, G. M., Molenberghs, G., & Lipsitz, S. R. (1995). Regression-models for longitudinal binary responses with informative drop-outs. *J. R. Stat. Soc. Ser. B-Methodol.*, **57**(4), 691–704.

Gange, S. J. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm. *Am. Stat.*, **49**(2), 134–138.

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. *Stat. Model.*, **1**, 81–102.

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

Hin, L. Y., Carey, V. J., & Wang, Y. G. (2007). Criteria for working-correlation-structure selection in gee: Assessment via simulation. *Am. Stat.*, **61**(4), 360–364.

Hin, L. Y. & Wang, Y. G. (2009). Working-correlation-structure identification in generalized estimating equations. *Stat. Med.*, **28**(4), 642–658.

Horton, N. J., Bebchuk, J. D., Jones, C. L., Lipsitz, S. R., Catalano, P. J., Zahner, G. E. P., & Fitzmaurice, G. M. (1999). Goodness-of-fit for gee: An example with mental health service utilization. *Stat. Med.*, **18**(2), 213–222.

Hosmer, D. W., Hosmer, T., leCessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Stat. Med.*, **16**(9), 965–980.

Hosmer, D. W. & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression-model. *Comm. Statist. Theory Methods*, **9**(10), 1043–1069.

Jokinen, J. (2006). Fast estimation algorithm for likelihood-based analysis of repeated categorical responses. *Comput. Stat. Data Anal.*, **51**(3), 1509–1522.

Lang, J. B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Ann. Stat.*, **24**(2), 726–752.

Lang, J. B. (2005). Homogeneous linear predictor models for contingency tables. *J. Am. Stat. Assoc.*, **100**(469), 121–134.

Lang, J. B. & Agresti, A. (1994). Simultaneously modelling joint and marginal distributions of multivariate categorical responses. *J. Am. Stat. Assoc.*, **89**(426), 625–632.

Lee, A. J. (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association. *Am. Stat.*, **47**(3), 209–215.

Lee, J. H. & Qaqish, B. F. (2004). Modified gee and goodness of the marginal fit (gomf) test with correlated binary responses for contingency tables. *Biom. J.*, **46**(6), 675–686.

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. *J. R. Stat. Soc. Ser. B-Methodol.*, **54**(1), 3–40.

Liu, I., Mukherjee, B., Suesse, T., Sparrow, D., & Park, K. P. (2009). Graphical diagnostics to check model misspecification for the proportional odds regression model. *Stat. Med.*, **28**, 412–429.

Liu, I. & Suesse, T. (2008). The analysis of stratified multiple responses. *Biom. J.*, **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.* New York: Chapman and Hall, 2nd edition.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Stat. Assoc.*, **92**(437), 162–170.

Molenberghs, G. & Verbeke, G. (2004). Meaningful statistical model formulations for repeated measures. *Stat. Sin.*, **14**(3), 989–1020.

Pan, W. (2001a). Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**(1), 120–125.

Pan, W. (2001b). Model selection in estimating equations. *Biometrics*, **57**(2), 529–534.

Pan, W. (2002). Goodness-of-fit tests for gee with correlated binary data. *Scand. J. Stat.*, **29**(1), 101–110.

Pan, W. & Connett, J. E. (2002). Selecting the working correlation structure in generalized estimating equations with application to the lung health study. *Stat. Sin.*, **12**(2), 475–490.

Pan, W., Connett, J. E., Porzio, G. C., & Weisberg, S. (2001). Graphical model checking with correlated response data. *Stat. Med.*, **20**(19), 2935–2949.

Pan, W. & Le, C. T. (2001). Bootstrap model selection in generalized linear models. *J. Agric. Biol. Environ. Stat.*, **6**(1), 49–61.

Preisser, J. S. & Qaqish, B. F. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika*, **83**(3), 551–562.

R-Development-Core-Team (2006). R: A language and environment for statistical computing.

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. *J. Comput. Graph. Stat.*, **9**(1), 141–157.

Rotnitzky, A. & Jewell, N. P. (1990). Hypothesis-testing of regression parameters in semiparametric generalized linear-models for cluster correlated data. *Biometrika*, **77**(3), 485–497.

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.

Wang, Y. G. & Hin, L. Y. (2009). Modeling strategies in longitudinal data analysis: Covariate, variance function and correlation structure selection. *Comput. Stat. Data Anal.*, **54**(12), 3359–3370.

Ware, J., Snow, K., Kosinski, M., & Gandek, B. (2000). Sf-36 health survey: Manual and interpretation guide.

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

Wooden, M., Freidin, S., & Watson, N. (2002). The household, income and labour dynamics in australia (hilda) survey: wave 1 survey methodology. *Australian Econ. Rev.*, **35**(3), 339348.

Yan, J. & Fine, J. (2004). Estimating equations for association structures. *Stat. Med.*, **23**(6), 859–874.

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.