Modelling Strategies for Repeated Multiple Response Data

Thomas Suesse[1] and Ivy Liu[2]
[1] *Centre for Statistical and Survey Methodology*
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

This article discusses modelling strategies for repeated measurements of a multiple response variable. This refers to a categorical variable where one can select none or more than one of the categories. We consider each of the response categories as a binary response and model the means using a marginal model approach. A generalised estimating equations (GEE) method is used to account for different correlation structures, both over time and between items. In addition, we discuss an alternative approach using a generalised linear mixed model (GLMM) with conditional interpretations. The GLMM contrasts with the marginal model which shows population-averaged interpretations. We illustrate these methods using The Household, Income and Labour Dynamics in Australia (HILDA) Survey, a household-based panel study.

Key words: Repeated Measurements, Generalised Linear Models (GLMs), Generalised Estimating Equations (GEE), Generalised Linear Mixed Models (GLMMs)

1 Introduction

Surveys often contain categorical responses where respondents may select none or more than one responses. This occurs, for example, when respondents are asked to "tick all that apply" for a list of the outcome categories. These are referred to as multiple response data in this paper. As an example, consider the Household, Income and Labour Dynamics in Australia (HILDA) Survey. It is a longitudinal survey that collects information about economic and subjective well-being, labor market dynamics and family dynamics, beginning in 2001. For annual expenses, respondents are asked to "tick all that apply" for various categories such as holidays and holiday travel costs, private health insurance, other insurance (such as home and contents and motor vehicles), etc. Following the literature, we refer to each category as an item.

The analysis of multiple responses has been considered by various authors. Agresti & Liu (1999, 2001) discussed different modeling strategies to describe the association between items and explanatory variables, by treating 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. Loughin & Scherer (1998) developed methods to test for the independence between each of the response items and an explanatory variable. Bilder & Loughin (2002) considered the case where the data are stratified by a third variable, and developed a test to detect whether the group and items are marginally independent given the stratification variable. Bilder & Loughin (2004) developed a test for marginal independence between two categorical variables with multiple responses. Bilder & Loughin (2007) showed the case of modelling two or more categorical variables across all items simultaneously using loglinear models. Liu & Suesse (2008) presented two methods for making inference for multiple responses when the data include highly stratified variables.

However, the situations with repeated measurements or longitudinal data have not been considered by these authors. The approach proposed by Bilder & Loughin (2007) could be applied to repeated multiple response data, by treating the two variables as repeated

multiple responses, but it leaves out the distinct correlation structure over time. Another possible approach uses hierarchical models where the responses of being selected or not for each of the items are grouped at two different levels across items and time points. Because the hierarchical models use various variance components to take into account the dependency among items and time points, it does not directly incorporate correlation structures between repeated measurements. The hierarchical model is a special case of a generalised linear mixed model (GLMM), which will be discussed in Section 3.

Extending the modeling strategies given by Agresti & Liu (1999, 2001), this paper discusses methodologies for repeated multiple responses by considering suitable association structures across both items and 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 shows methods based on the marginal model approach. We focus primarily on generalised estimating equations or GEE (Liang & Zeger, 1986) and consider a variety of possible correlation structures. Standard GEE methods allow only a few options for the correlation structure, and these options are unlikely to be valid for repeated multiple response data. Due to the dependency across several levels, we propose an alternative method combining the levels in a way that allows the use of standard GEE methods while also accounting for multiple correlated levels. 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. It confirms the usefulness of the proposed method. Section 5 illustrates the methods on the HILDA survey using waves E, F, G and H (years 2005-2008). We also use some of the goodness-of-fit techniques for GEE methods to evaluate the choice of the correlation structure. The paper ends with a discussion.

2 Marginal Modelling

Agresti & Liu (2001) considered several strategies to model c items simultaneously for a single time point. They introduced the marginal model approach that takes the dependence between items on the same subject into account. Generalising their method, we first consider the marginal model approach to model c items and T time points simultaneously.

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. 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. The ith subject's $2^{c \cdot T}$ response profile for c items and T time points becomes $\mathbf{Y}_i = [(Y_{i11}, Y_{i21}, \ldots, Y_{ic1}), \ldots, (Y_{i1T}, Y_{i2T}, \ldots, Y_{icT})]'$ with the mean $\boldsymbol{\pi}_i = [(\pi_{1|i1}, \pi_{2|i1}, \ldots, \pi_{c|i1}), \ldots, (\pi_{1|iT}, \pi_{2|iT}, \ldots, \pi_{c|iT})]'$.

Using a logit link $h(\pi) = \text{logit}(\pi) = \log(\pi/(1-\pi))$, a marginal model can be written as

$$h(\boldsymbol{\pi}_i) = \mathbf{X}_i \boldsymbol{\beta}, \tag{1}$$

where $h(\boldsymbol{\pi}_i) = [(h(\pi_{1|i1}), h(\pi_{2|i1}), \ldots, h(\pi_{c|i1})), \ldots (h(\pi_{1|iT}), h(\pi_{2|iT}), \ldots, h(\pi_{c|iT}))]'$ and $\mathbf{X}_i = [(\mathbf{x}_{i11}, \ldots, \mathbf{x}_{ic1}), \ldots, (\mathbf{x}_{i1T}, \ldots, \mathbf{x}_{icT})]'$ is the design matrix associated with covariate vector $\mathbf{x}_{ijt}$ of the ith subject, the jth item and tth time point, and $\boldsymbol{\beta}$ is the vector of population-averaged effects. They are also often referred to as fixed effects, marginal or mean model parameters, because model (1) refers to the first moments. In general, $h(\cdot)$ can be any smooth link function. Other popular choices are the log and probit link.

For repeated multiple responses, $\{Y_{ijt}\}$ are correlated over items (j) as well as over the time points (t). There are several ways of modelling the correlation across items and time. A naive approach is to apply a generalised linear model (GLM) (McCullagh & Nelder, 1989) by assuming independent responses, however this approach does not yield proper standard errors for $\hat{\boldsymbol{\beta}}$. We next discuss two methods to incorporate the correlation – maximum likelihood and GEE.

2.1 Maximum Likelihood method

Model (1) can be expressed as a generalised log-linear model and one can use maximum likelihood (ML) to estimate the parameters (Lang & Agresti, 1994; Lang, 1996). Lang (2005) proposed an extension that allows any smooth link function $h(\cdot)$.

The ML method treats the counts from the $2^{c \cdot T}$ response profile for each of the K covariate settings as a multinomial distribution. In general, we can across-classify subjects according to their response profile and covariates to create a $K \times 2^{c \cdot T}$ contingency table. ML estimates are obtained by maximizing the multinomial likelihood subject to constraints satisfying the mean model. The method requires that the contingency table has large cell counts in order to apply the central limit theorem. When the number $2^{c \cdot T}$ or K is very large, the requirement is not fulfilled. Continuous covariates are of particular concern, because then K often equals the sample size. Another similar application where the sparseness of the data causes problems for the ML method has been illustrated by Suesse & Liu (2012).

Additionally, one could also model the second order moments through the odds ratio

$$\theta_{i,jj',tt'} = \frac{\Pr(Y_{ijt} = 1, Y_{ij't'} = 1) \Pr(Y_{ijt} = 0, Y_{ij't'} = 0)}{\Pr(Y_{ijt} = 1, Y_{ij't'} = 0) \Pr(Y_{ijt} = 0, Y_{ij't'} = 1)}, j \neq j' \text{ or } t \neq t' \tag{2}$$

or the correlation

$$R_{i,jj',tt'} = \frac{\Pr(Y_{ijt} = 1, Y_{ij't'} = 1) - \Pr(Y_{ijt} = 1) \Pr(Y_{ij't'} = 1)}{[\Pr(Y_{ijt} = 1)(1 - \Pr(Y_{ijt} = 1)) \Pr(Y_{ij't'} = 1)(1 - \Pr(Y_{ij't'} = 1))]^{1/2}}. \tag{3}$$

Both approaches lead to a more complicated ML, because it requires fitting the mean and association models jointly. Fitzmaurice & Laird (1993) proposed using the conditional odds ratio to model the association, leading to a complicated ML approach that utilises the iterative proportional fitting algorithm.

Another measure of association is the dependence ratio (Ekholm et al., 1995, 2003). In contrast to the correlation and the odds ratio, the dependence ratio is a linear function of the $K \times 2^{c \cdot T}$ multinomial probabilities, reducing the complexity of the ML method. However

it does not eliminate the problems associated with the large cell count requirement. The R-package (R-Development-Core-Team, 2006) `drm` (Jokinen, 2009) can be used to fit such models. The dependence ratio was also criticised by Molenberghs & Verbeke (2004), who argued instead for the odds ratio as a measure of association, due to symmetry and ease of interpretation.

2.2 Generalised Estimating Equations

Besides the ML method, the GEE method (Liang & Zeger, 1986) is another popular fitting procedure and is an extension of the quasi-likelihood method (Wedderburn, 1974) for multivariate data. The GEE method fits the marginal model (1) simultaneously across items and time points, while it also incorporates a chosen correlation structure, known as the *working correlation*. Let the working correlation structure be $\mathbf{R}_i(\boldsymbol{\alpha})$, a $cT \times cT$ correlation matrix for subject i $(i = 1, \ldots, n)$ depending on correlation parameter(s) $\boldsymbol{\alpha}$. The GEE estimates are obtained by computing the root of the generalised estimation equations

$$\sum_{i=1}^{n} \mathbf{M}_i' \mathbf{V}_i^{-1} \mathbf{r}_i(\boldsymbol{\beta}) = \mathbf{0}, \tag{4}$$

where $\mathbf{M}_i = \partial \boldsymbol{\pi}_i / \partial \boldsymbol{\beta}$, $\mathbf{V}_i = \mathbf{A}_i \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i$, $\mathbf{r}_i(\boldsymbol{\beta}) = \mathbf{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta})$ and matrix $\mathbf{A}_i = \text{Diag}(\sqrt{\text{Var}(\mathbf{Y}_i)})$ with $\text{Var}(\mathbf{Y}_i) = \boldsymbol{\pi}_i(1 - \boldsymbol{\pi}_i)$, because we are dealing with binary observations. Preisser & Qaqish (1996) suggested the iterated weighted least squares method to obtain $\widehat{\boldsymbol{\beta}}$. The (co)variance estimator for $\widehat{\boldsymbol{\beta}}$ is $(\sum_{i=1}^{n} \widehat{\mathbf{M}}_i' \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{M}}_i)^{-1}$, called the naive (co)variance. One can adjust the standard errors for $\widehat{\boldsymbol{\beta}}$ to reflect what actually occurs for the sample data by using the 'sandwich' estimator, also known as the robust variance (Liang & Zeger, 1986). 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. Note that the robustness of the robust variance only applies to a mis-specification of

the association model, but not to mis-specification of the clusters. For example, for repeated multiple responses, if a cluster is wrongly specified as a sequence of less than $c \cdot T$ binary responses, then the robust variance is not consistent.

Choosing a suitable correlation model/structure, one that is close to the true model, is essential in obtaining more efficient parameter estimates for $\boldsymbol{\beta}$ (Liang & Zeger, 1986). Considering that items are often either very similar or different, suitable correlation structures between items j_1 and j_2 include 'independence' ($R_{j_1 j_2} = 0$), 'exchangeable' ($R_{j_1 j_2} = \alpha$), and 'unstructured' ($R_{j_1 j_2} = \alpha_{j_1, j_2}$), see Agresti & Liu (1999, 2001). For longitudinal data, 'autoregressive' (AR) ($R_{t_1 t_2} = \alpha^{|t_1 - t_2|}$) is the common one, because observations further apart in time are generally less correlated than those closer in time. The autoregressive structure can be generalised into the m-dependence structure, where the responses are not correlated if the time lag exceeds m units.

2.3 Alternating Logistic Regression and Method of Orthogonalised Residuals

Alternatively, rather than using the correlation as the association, one can use the pairwise odds ratio $\theta_{i,jj',tt'}$ instead, see (2). A typical association model has the form

$$g(\theta_{i,jj',tt'}) = \mathbf{Z}_{i,jj',tt'}\boldsymbol{\alpha} \tag{5}$$

where $g(\cdot)$ is a link function, $\boldsymbol{\alpha}$ is a vector of model parameters and $\mathbf{Z}_{i,jj',tt'}$ is a row vector of predictors. Fitting (1) and (5) jointly was suggested by Lipsitz et al. (1991) using the standard GEE method for both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. For a popular choice of $h(\cdot) = \text{logit}(\cdot)$ and $g(\cdot) = \log(\cdot)$, Carey et al. (1993) proposed the alternating logistic regression (ALR) method, where two estimating equations of similar form as those for logistic regression are iterated in an alternating pattern. However, these two estimating equations of ALR are not orthogonal, meaning that mis-specification of the association model leads possibly to inconsistent $\hat{\boldsymbol{\beta}}$. To overcome this limitation, Zink (2003) extended ALR by adding an additional parameter that

makes residuals and responses asymptotically orthogonal. This method is implemented in the R-package `orth` (By et al., 2011) with ALR as a special case.

Similar to the correlation structure, the options 'exchangeable' and 'unstructured' are available for the pairwise odds ratios by specifying the design matrix $\mathbf{Z}_{i,jj',tt'}$. As an analogue to the autoregressive correlation structure, Molenberghs & Verbeke (2005) suggested the model $\log(\theta_{i,jj,t_1t_2}) = \frac{1}{|t_2-t_1|}\alpha$ or equivalently $\theta_{i,jj,t_1t_2} = \exp(\alpha)^{\frac{1}{|t_2-t_1|}}$. The θ value indicates the direction of association between responses, where '1' indicates independence, '< 1' indicates a negative association, and '> 1' corresponds to a positive association. The function $\exp(\alpha)^{\frac{1}{|t_2-t_1|}}$ guarantees that the association diminishes (approaches 1) as $|t_2 - t_1|$ increases.

2.4 Combining Levels for Repeated Multiple Response Data

Repeated multiple responses are characterised by both items and time points For a common time point, the chosen association model between responses should match the structure for standard multiple responses. Similarly, for a common item, the association model used across two time points should be one for longitudinal data. The question of how to combine the two levels effectively is considered next.

2.4.1 Combining Levels for GEE

Let $R_{j_1j_2,t_1t_2}$ denote the correlation for two responses $Y_{ij_1t_1}$ and $Y_{ij_2t_2}$. When either $j_1 = j_2$ or $t_1 = t_2$, the correlation should match the marginal correlation structure, that is, $R_{j_1j_2,t_1t_1} = R_{j_1j_2}$ or $R_{j_1j_1,t_1t_2} = R_{t_1t_2}$. The $R_{j_1j_2} \in (-1, 1)$ stands for the correlation between two items with common time point, and the $R_{t_1t_2}$ stands for the correlation between two time points within the same item, which is usually non-negative. In general, we let the correlation between two time points t_1 and t_2 vary for different items j. The underlying assumption is that the correlation between items does not change over time (which is reasonable for a small number of time points but might need to be adapted for many time points) and that the correlation over time might differ for different items. When $j_1 \neq j_2$ and $t_1 \neq t_2$, we suggest

two options

$$R_{j_1j_2,t_1t_2} = \begin{cases} \sqrt{R_{j_1j_1t_1t_2,}} \times \sqrt{R_{j_2j_2,t_1t_2}} \times R_{j_1j_2} & \text{Option 1} \\ 0 & \text{Option 2} \end{cases} \tag{6}$$

If $R_{t_1t_2} := R_{j_1j_1,t_1t_2} = R_{j_2j_2,t_1t_2}$ for any two items j_1 and j_2, then Option 1 becomes $R_{j_1j_2,t_1t_2} = R_{j_1j_2} \times R_{t_1t_2}$.

The motivation for using the 'product' in Option 1 is based on the following desirable properties: i) $R_{j_1j_2,t_1t_2} = R_{j_1j_2,t_1+k,t_2+k}$, if $R_{jj,t_1t_2} = R_{jj,t_1+k,t_2+k}$; ii) $R_{j_1j_2,t_1t_2}$ diminishes if $|t_2 - t_1|$ increases, provided the same applies to $R_{j_1j_1,t_1t_2}$ and $R_{j_2j_2,t_1t_2}$; iii)$R_{j_1j_2,t_1t_2} \leq 0$ iff $R_{j_1j_2} \leq 0$ assuming $R_{jj,t_1,t_2} \geq 0$, and iv) $R_{j_1j_2,t_1t_2} \leq \min(\sqrt{R_{j_1j_1,t_1t_2} \times R_{j_2j_2,t_1t_2}}, R_{j_1,j_2})$. Option 2 refers to independence across different items over different time points.

Unfortunately, when the AR structure is used for the time dependence in model (6), we cannot apply standard statistical packages, such as `gee` and `geepack` in `R`, to fit the proposed models. Besides, because AR is a non-linear correlation model, we cannot specify a design matrix for a linear correlation model using the package `geepack`. To solve the complication, we propose another feasible option using existing software packages to incorporate model (6).

In step 1, we fit the mean model by choosing the AR structure for the time dependence and ignore item dependence. In Step 2, we fit the same mean model again but this time use an appropriate structure for the items and ignore time dependence.

The mean model parameters $\boldsymbol{\beta}$ are estimated consistently provided the mean model is correctly specified, irrespective of the association model. The corresponding correlation parameters $\boldsymbol{\alpha}$ in Step 1 and 2 are estimated also consistently, provided those models for items and time points are correct, but irrespectively of the correlation model for those items for which $j_1 \neq j_2$ and $t_1 \neq t_2$. This is case, because the two sets of residuals used in Step 1 and 2, over which the correlation model parameters are estimated, are distinct sets. This applies for the standard GEE method only (sometimes called GEE1), for which the estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are orthogonal but not for GEE2 (Liang et al., 1992).

In step 3, we use the estimates from step 1 and 2 to compute the fixed working corre-

lation $\mathbf{R}_i$ following Options 1 or 2 in (6), and re-fit the mean model again with the 'fixed' working correlation structure $\mathbf{R}_i$. The option 'fixed' is standard for most GEE packages. To investigate the performance of this 3-step method, a simulation study is conducted in Section 4.

2.4.2 Combining Levels for ALR

We propose the following odds ratio model for repeated multiple response data:

$$\theta_{j_1,j_2,t_1,t_2} = \begin{cases} \exp(\alpha_{j1,j2})^{\frac{1}{|t_2-t_1|+1}} & \text{for } j_1 \neq j_2 \text{ (different items)} \\ \exp(\alpha_{j_1})^{\frac{1}{|t_2-t_1|}} & \text{for } t_1 \neq t_2 \text{ and } j_1 = j_2 \text{ (equal items)} \end{cases} \tag{7}$$

This model has similar properties as the proposed correlation model, see (6), and it extends the idea behind the model suggested by Molenberghs & Verbeke (2005) for longitudinal data. Notice that, the exponent for $j_1 \neq j_2$ is $|t_2 - t_1| + 1$ instead of $|t_2 - t_1|$ to allow $t_1 = t_2$.

2.5 Model Diagnostics

GEE requires to specify a mean model, such as equation (1), and a model for the association. Since GEE is not a likelihood based fitting approach, likelihood-based inference is not possible. There is a vast literature on model diagnostics for GEE, including ALR as a sub-case. For example, when the focus is primarily on efficiency of $\hat{\boldsymbol{\beta}}$, Pan & Connett (2002) considered several methods for choosing a working correlation subject to minimizing the predictive mean squared error. Other examples include diagnostics to check the association model (Rotnitzky & Jewell, 1990; Hin et al., 2007) and the overall goodness-of-fit (Barnhart & Williamson, 1998; Horton et al., 1999; Pan, 2001) following the principle of the famous Hosmer and Lemeshow statistic. Some of these diagnostics are illustrated in Section 5 for the HILDA data set.

3 Generalised Linear Mixed Models

The marginal model (1) 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 (GLMMs) include random or subject-specific effects in the mean model, additional to the fixed effects. This model is referred to as a subject-specific model, since parameters are interpreted 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. Define $\tilde{\boldsymbol{\pi}}_i = E(\mathbf{Y}_i|\mathbf{u}_i)$. Given the random effects $\mathbf{u}_i$ for cluster i, the GLMM has a similar form as a GLM

$$h(\tilde{\boldsymbol{\pi}}_i) = \mathbf{X}_i\boldsymbol{\beta}^{sub} + \mathbf{Z}_i\mathbf{u}_i, \tag{8}$$

where the design matrix $\mathbf{Z}_i$ consists of rows $\mathbf{z}'_{ijt}$ referring to subject i, item j and time point t. The vector of fixed effects of the subject-specific model is $\boldsymbol{\beta}^{sub}$ to distinguish it from $\boldsymbol{\beta}$ of the marginal model.

It is commonly assumed that the random effects $\mathbf{u}_i$ of dimension r ($r \leq c \times T$) follow a multivariate normal distribution with mean $\mathbf{0}$ and unknown covariance matrix $\boldsymbol{\Sigma}$. Maximising the marginal likelihood that integrates out the random effects results in ML estimates for $\boldsymbol{\beta}^{sub}$ and $\boldsymbol{\Sigma}$ (Agresti, 2002). The integral usually cannot be solved analytically and numerical methods (such as Gauss-Hermite quadrature) must be applied. They work well for small r, 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 (Stiratelli et al., 1984; Schall, 1991; Breslow & Clayton, 1993; Zeger et al., 1988; Goldstein, 1991; Raudenbush et al., 2000). However, most of them can yield poor estimates, in particular for first order expansions (Breslow & Lin, 1995). 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 by treating the random effects as unobserved data. Algorithms have been provided by McCulloch (1997) and Booth & Hobert (1999) among others.

A special case of GLMMs includes a typical multilevel approach that considers subjects, time and items as levels (Goldstein, 1991; Guo & Zhao, 2000), i.e. $u_j \sim N(0, \sigma^2_{j,item})$ (item effects), $u_t \sim N(0, \sigma^2_{t,time})$ (time effects) and $u_i \sim N(0, \sigma^2_{cluster})$ (cluster effects). This approach does not resemble a typical time dependence structure, such as the autoregressive structure. A simple case of such a multi-level model uses random intercepts only, 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, the covariance is 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 a GLMM is not useful in modelling data with negative correlations. This is supported by the simulation study in the next section, where under a marginal model with negative correlation the type I error rates exceed 5% when a GLMM was fitted.

There are often exact or approximate relationships between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^{sub}$, depending on $h(\cdot)$ (Zeger et al., 1988). For example for the logit link

$$\mathbf{x}'_{ijt}\boldsymbol{\beta} \approx a(\boldsymbol{\Sigma})\mathbf{x}'_{ijt}\boldsymbol{\beta}^{sub}, \tag{9}$$

where $a(\boldsymbol{\Sigma})$ is a constant depending on the random effects variance and on $\mathbf{z}'_{ijt}$. It illustrates that a mixed model and a marginal model are different. Detailed arguments for the choice between marginal and subject-specific models are available in Agresti (2002) and Neuhaus (1992). Often and particularly in surveys, where interpretation is sought for the whole population, marginal models are more relevant.

4 Simulation Study for GEE Approach Combining Levels

To investigate the performance of our proposed 3-step method, we consider $c = 3$ and $T = 3$ in a simulation study. This choice of c and T is reasonable to illustrate the working correlation 'unstructured' for the items and the AR-structure over the time points.

Marginal Mean Models

We consider two marginal mean models. Model A has the form

$$\text{logit}(\pi_{j|it}) = \beta_{0j} + \beta_{1j}X$$

with $\beta_{01} = 0.0$, $\beta_{02} = 0.5$,$\beta_{03} = 0.9$, $\beta_{11} = 0.0$, $\beta_{12} = 0.5$, $\beta_{13} = 1.0$ and $X \sim N(0, 1)$. The intercept β_{01} and slope β_{11} for the first item are both set to zero in order to investigate the type I error rates.

Model B has the form

$$\text{logit}(\pi_{j|it}) = \beta_0 + \beta_1 X,$$

where $\beta_0 = -1$ and $\beta_1 = 3$ with $X \sim N(0, 1)$, with no item or time effects. The number of clusters (or subjects) generated for each model is $n = 30, 100, 500$.

Association Models

We consider three association models. For simplicity we assume for the first two models a common AR structure $R_{jj,t_1t_2} = R_{t_1t_2} = 0.3^{|t_2-t_1|}$ and specify the between item correlation $(R_{j_1j_2})$ as $R_{1,2} = 0.2$, $R_{1,3} = 0.1$ and $R_{2,3} = 0.3$. Model I reflects Option 1 in (6). Option 2 was also included in simulation study but results are not shown. Model II refers to $R_{j_1j_2,t_1t_2} = 0.05$ for $j_1 \neq j_2$ and $t_1 \neq t_2$. Model III uses the odds ratio association in (7) with $\theta_{1,2,t_1,t_2} = 0.3^{\frac{1}{|t_2-t_1|+1}}$, $\theta_{1,3,t_1,t_2} = 0.4^{\frac{1}{|t_2-t_1|+1}}$, $\theta_{2,3,t_1,t_2} = 0.5^{\frac{1}{|t_2-t_1|+1}}$, and $\theta_{j,j,t_1,t_2} = 5^{\frac{1}{|t_2-t_1|}}$. Notice that model III imposes negative correlations between the 3 items, because $\theta < 1$.

Data Generation from Multivariate Binary

To simulate the data, we need to calculate the joint distribution for each $\mathbf{Y}_i$ specified by $2^{c \cdot T}$ probabilities, for the given marginal means $\pi_{j|it}$ and correlations. Note that when Model III is used, the correlations can be computed from the marginal probabilities and the odds ratio association. From the correlations and $\pi_{j|it}$, the pair-wise probabilities $\Pr(Y_{ijt} = 1, Y_{ij't'} = 1)$ are computed. Finally a joint distribution can be found, subject to $\Pr(Y_{ijt} = 1, Y_{ij't'} = 1)$ and $\pi_{j|it}$ (Lee, 1993). There are usually many solutions, provided a feasible solution exists. The iterative proportional fitting algorithm (IPF) of Gange (1995) is applied to obtain such a solution, which is analogous to the simulation study in Bilder et al. (2000).

Mixed Models

We also generate data from two mixed models, denoted by model A* and B* with the same fixed effect parameters as for models A and B. Let the item random effects be $u_{ij}^{item} \sim N(0, 0.5)$, time random effects be $u_{it}^{time} \sim N(0, 1)$ and the cluster random effects be $u_i^{cluster} \sim N(0, 0.3)$ for $j = 1, 2, 3$, $t = 1, 2, 3$ and $i = 1, \ldots, n$. All random effects are assumed as independent.

The fixed effects of a GLMM and of a marginal model have different meanings and estimates are of different magnitude. From the approximate relationship (9) follows that approximately $\hat{\boldsymbol{\beta}} = 0$ iff $\hat{\boldsymbol{\beta}}^{sub} = 0$. Therefore testing for significance of a covariate can be achieved via the GLMM approach or the marginal model approach. To assess the type I errors under either model approach, we included both approaches into the simulation study.

Fitting Methods

Under the marginal models A and B and the mixed model A* and B*, we fit the marginal model with several choices of the association model, see *a*) - *k*) below, and also fit a mixed model, see *l*). In both cases, we correctly specify the (conditional) mean model. The fitting methods considered are:

a) unstructured for whole cluster (unstr)

b) exchangeable for whole cluster (exch)

c) independence for whole cluster (ind)

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

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

f) Option 1 specified by (6), mean model and working correlation estimated jointly (opt1-j)

g) Option 1, item and time correlation estimated separately, i.e. 3-step method applied (opt1-s)

h) same as f), but Option 2 (opt2-j)

i) same as g), but Option 2 (opt2-s)

j) ALR specified by (7) denoted by (alr1)

k) same as j) except $\theta_{j_1,j_2,t_1,t_2} = 1$ for $j_1 \neq j_2$ and $t_1 \neq t_2$ (alr2)

l) mixed model with item, time and subject random effects (glmm)

Results of the simulation study are shown in Tables 2, 3, 4, 5 and 6. For $n = 30$ and $n = 100$, we simulated $10,000$ data sets, and for $n = 500$, we simulated only $5,000$ data sets. The effect of mis-specification of the association model on the estimation of the fixed effects can be assessed.

Tables 2 (for Model A) and 3 (for Model B) show the relative mean squared error (RMSE), which is the mean squared error relative to the method using the correct known (fixed) correlation structure to evaluate the relative efficiency. The tables also give the coverage for a 95% confidence interval based on the naive variance and on the robust variance (denoted by "naive" and "robust"). The information on the bias of standard errors can be followed,

because under-coverage indicates that standard errors are under-estimated and vice versa. Models A and B contain several parameters. Because results are similar for all parameters and to summarise the results more effectively, the tables show only a single value for each model. The single value shows the average of RMSEs (or coverages) over all fixed effects.

Table 4 and Table 5 respectively show the type I error for β_{01} and β_{11} in models A and A*. The type I error is computed as the proportion of times that the null hypothesis (such as $\beta_{01} = 0$) was rejected based on the Wald test at the 5% significance level. We can assess whether GEE and GLMMs are robust under mis-specification by assessing the type I error. Models I and II, as well as the GLMM, refer to positive correlations between items, whereas model III refers to negative. Table 6 shows the mean squared error (MSE), "naive", and "robust" for each of the two parameters in model B. It allows us to compare the results of each parameter with the average ones in Tables 2 and 3.

Before we draw conclusions from the tables, there are two notices. Firstly, these tables do not show the confidence interval (CI) length and the bias. The bias is negligible. The CI length is monotone in the coverage, because the CI is centered around the same $\hat{\boldsymbol{\beta}}$ due to the consistency of $\hat{\boldsymbol{\beta}}$. Therefore, in general, a method with a smaller coverage has a shorter CI. Secondly, all methods *a*) - *k*) were applied to the whole cluster of size $c \cdot T$ except methods *d*) and *e*). Methods *d*) and *e*) 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.

Figures 1, 2 and 3 present the results graphically. For better scaling in Figure 1 the RMSE is the MSE relative to the best method.

In summary of Tables 2, 3, and 6, the larger the number of clusters becomes, the more accurate the robust variance due its consistency, except for methods 'time' and 'item'. The naive variance seems rather unreliable. The method 'unstr' usually performs poorly for a small n in terms of both relative efficiency and non-convergence, but improves when n increases. Methods 'opt2-j' and 'opt2-s', which assume zero correlation between responses of

the same person for different items and different time-points, are worse than methods 'opt1-j' and 'opt1-s'. Method 'opt1-j' seems generally best, but it requires users to write their own code to implement the method. The suggested and relatively easily implementable method 'opt1-s', as an alternative to method 'opt1-j', performs almost as well. We would expect a higher gain in efficiency if the correlation parameters and the fixed effects parameters are all estimated jointly, but surprisingly the gain in relative efficiency of 'opt1-j' compared to 'opt1-s' is almost negligible. Unexpectedly, method 'opt1-s' also preforms well under model III (odds ratio association models), which is probably due to a high non-convergence rate and some undesirable properties of ALR (Zink, 2003).

The Tables 4 and 5 show that GLMMs fail in maintaining the significance level under the marginal model. It is the worst for the association model III, which imposes negative correlations between item. It is not surprising, because GLMM fitting method do only provide a model based (naive) variance, but not a robust variance, as GEE (Zeger et al., 1988). GLMMs also impose a non-negative correlations between items, explaining its poor performance under model III. The tables also show that for a large n, GEE with the robust variance (except method 'item') can maintain the significance level even under the mixed model.

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. In the first wave (wave A, 2001) $13,969$ persons 15 years or older were successfully interviewed. Subsequent interviews for later waves were conducted about one year apart. The HILDA survey contains information on economic and subjective well-being, labour market dynamics and family dynamics collected through personal interviews and self-completion questionnaires. Details are

Figure 1: Averaged relative mean squared error of the GEE methods for mean models A and B with association models I and III

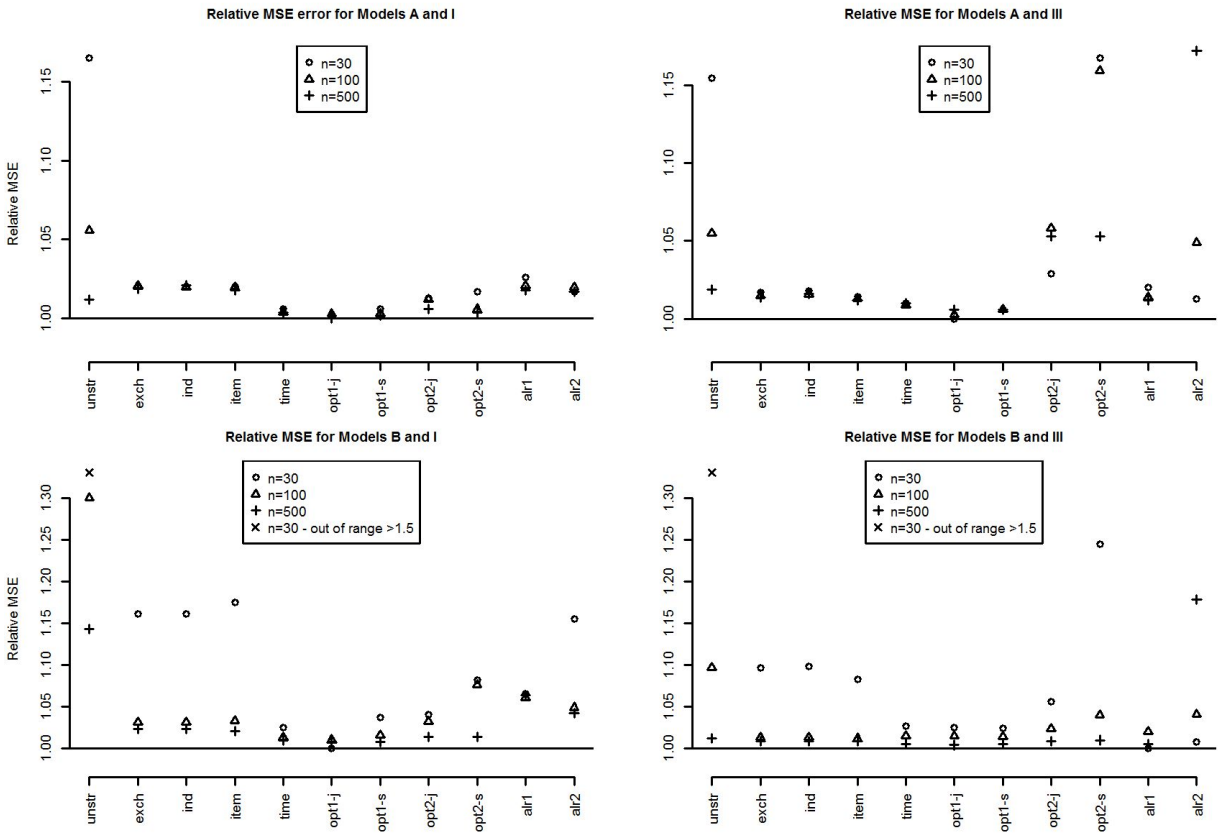


Figure 2: Averaged coverage of the GEE methods based on robust variance for mean models A and B with association models I and III

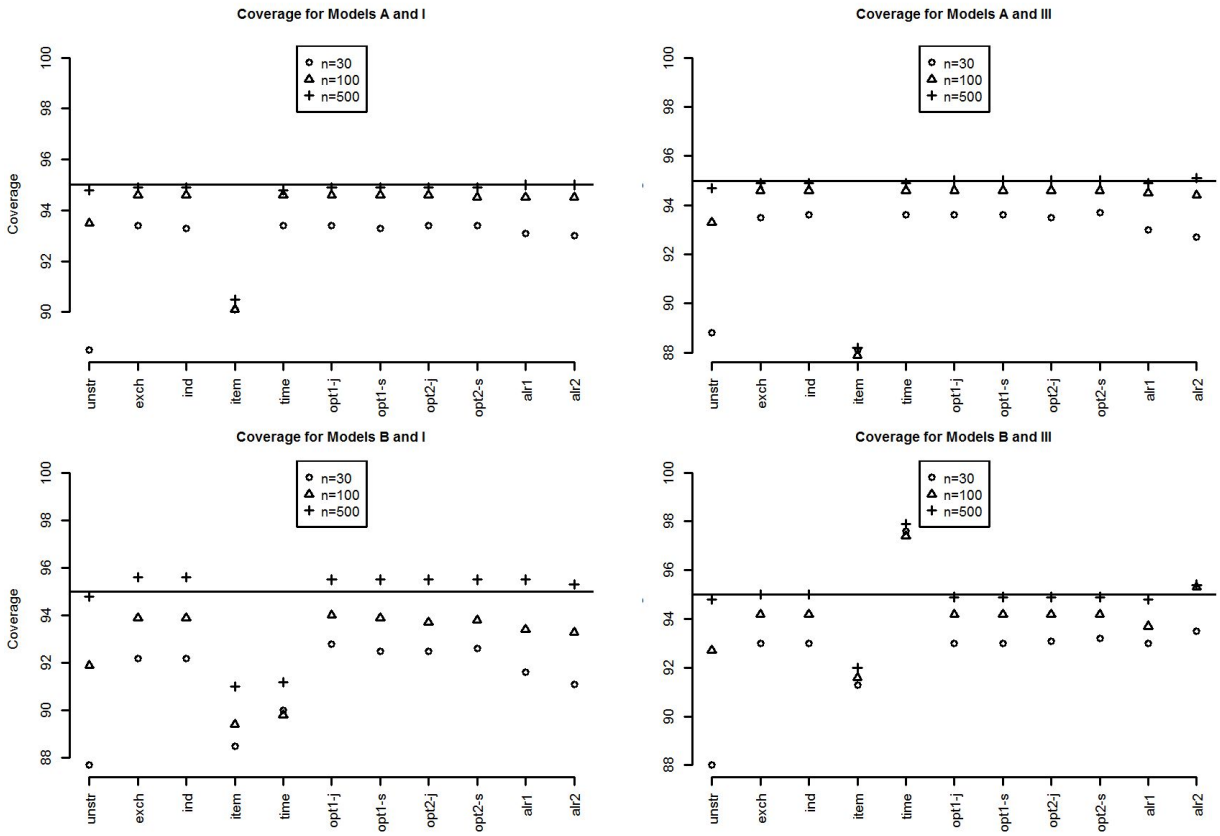
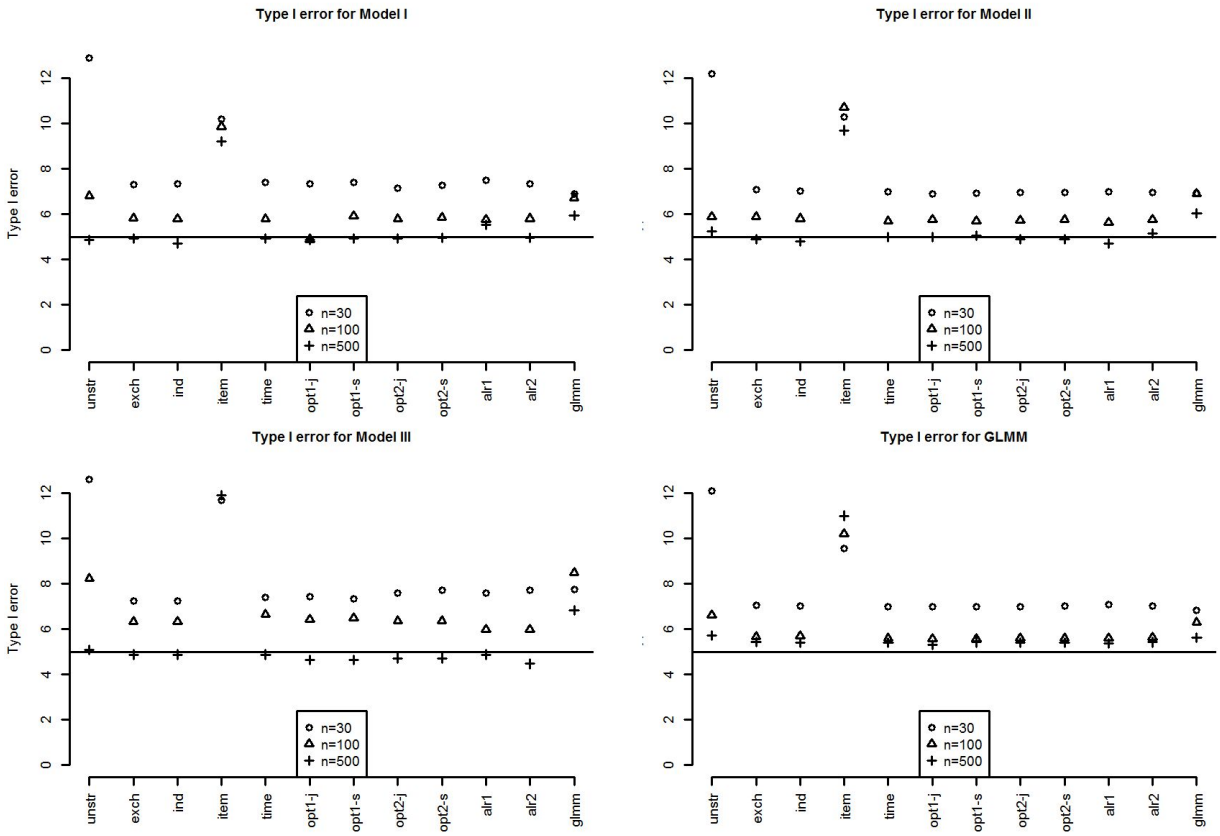


Figure 3: Type I error of the GEE methods and the GLMM method for the slope β_{11} for model A with association models I, II and III, and model A*



documented in Wooden et al. (2002).

In the self-completion questionnaire, respondents were asked about their daily, weekly, monthly and annual expenses. This paper focuses on three of the items for annual expenses: i) fees paid to doctors, dentists, opticians, physiotherapists, chiropractors and any other health practitioner (FD) (often referred to as 'extras'), ii) private health insurance (PHI), and iii) holidays and holiday travel costs (HOL).

In Australia, the government provides a compulsory basic health cover for everyone, called "Medicare", and purchasing a private health insurance as a top-up cover is optional. Therefore respondents might tick none, any one or two or all three of these three items. Item FD and HOL are available from wave E (2005) and the FD from wave F (2006). Therefore, we have 4 years data for two items and 3 years data for the first item. One of the research question governments and private health insurers might be interested in is how these three items relate to various covariates, such as sex, drinking, smoking, the long term health conditions, etc.

HILDA provides a number of health variables: i) alcohol drinking status (abstainer, ex-drinker, low risk, risky, high risk), ii) health scores (0-100) for mental health, general health, physical functioning and vitality from the SF-36 (Ware et al., 2000), iii) long term health conditions indicator (yes/no), developed at previous wave is denoted by 'developed T-1', at current wave denoted by 'developed T', iv) long term health conditions (e.g. speech/ hearing/ learning problems, limited use of feet/arms, shortness of breath, pain, mental health, etc.), v) smoking status (do not smoke, no longer smoke, smoke weekly but less often than daily, less often than weekly), vi) number of cigarettes a week, and vii) 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 additionally the household level. The notation $\pi_{j|ith}$ is the probability that item j was ticked at time t by subject i who was in household h. We need to use model diagnostics for GEE to select the best model. We first select the marginal model

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

by eliminating unnecessary covariates. Pan (2001) suggested a quasi-likelihood under the independence model criterion (QIC) for the GEE method. Because the standard AIC for GLMs is an approximation for QIC, we use AIC under the standard independence model assuming common fixed effects across all waves. The final mean model consists of $p = 156$ chosen covariates, 51 for FD, 47 for PHI and 58 for HOL.

Next, we need to assess the association model. As Pan (2001) noted, the above approximation of the QIC can only be used to check the mean model, not the working correlation. We also believe that the QIC is only useful for checking the mean model when the independence correlation is used and not for any other working correlation. Alternatively one might opt for the RJ (Rotnitzky & Jewell, 1990) criterion, where the working correlation with the smallest RJ should be chosen. Although Hin et al. (2007) noted that neither QIC nor the RJ criterion performed well in their simulation study for small datasets ($n = 100$ with cluster size 5), we expect the RJ criterion to perform well for large n, as for the HILDA survey.

GEE with the unstructured working correlation could not be fitted due to the large data set. We used the working correlation referring to Option 1 from (6), denoted by 'opt1' ($R_{t_1t_2} = R_{jj,t_1t_2} = R_{j'j',t_1t_2}$ for $j \neq j'$) and 'opt1*' ($R_{jj,t_1t_2} \neq R_{j'j',t_1t_2}$ for $j \neq j'$). Because the responses are correlated among three levels: items, time and household, Option 1 is extended by combining three levels, not only items and time. We also fitted the mean model

with various working correlations: independence (ind), exchangeable (ex), only accounting for time dependence (time), for items dependence (item) and for households dependence (HH). For ALR, we only show the results of models with time dependence (timeALR) and with time and household dependence (timeHH), because many options including (7) did not converge. Table 1 shows the results of the RJ criterion. According to the table, the best choice for GEE is 'opt1*' followed by 'time' and 'opt1' , but both ALR options also fit well.

Table 1: Assessing Working Correlation/Odds Ratio Models for HILDA

	Working Correlation							Working Odds Ratio	
Measure	opt1	opt1*	ind	ex	time	item	HH	time-ALR	timeHH
RJ	860	822	1822	1583	845	1907	1399	890	837

For 'opt1*', the AR parameters for the three items are 0.45 for FD, 0.90 for PHI and 0.54 for HOL. The correlation, between FD and PHI is -0.26, between FD and HOL is 0.21 and between PHI and HOL it is -0.16. The correlation between members of the same household is 0.53 for which items and time points are the same. For 'opt1' the single AR parameter is 0.70.

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 added a main effect for each of the major statistical regions and remoteness areas of Australia.

Finally to check the overall model fit for GEE with 'opt1*', we constructed goodness-of-fit (GOF) tests proposed by Horton et al. (1999) and Barnhart & Williamson (1998), extensions of the famous Hosmer & Lemeshow (1980) statistic. They used the idea of forming G groups by partitioning the space of covariates. With 10 groups, the test gave a p-value of 0.31. Thus, 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

$$\text{logit}(\pi_{j|ith}) = \mathbf{x}'_{ijth}\boldsymbol{\beta}_j^{sub} + u_h + u_{j|i} + u_{j|h} + u_{ti} + u_{th}, \tag{10}$$

assuming that these random effects are independent of each other. Instead of just accounting for a single random intercept effect, e.g. u_i, 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 models give more insight into the dependence of items across time-points and household members.

The fitting results for GEE ('opt1*'), ALR and GLMM are presented in Tables 7, 8 and 9. 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. This 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, but our primary focus of this paper is not on interpretation of such parameters. We emphasize on the statistical modeling and its influence on the associated p-values.

Parameter estimates for GEE, ALR 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. This can be explained by (9). The variance estimates of the random effects for u_h, $u_{j|i}$, $u_{j|h}$, u_{ti}, u_{th} are $\hat{\sigma}_h^2 = 6.34$, $\hat{\sigma}_{j|i}^2 = 3.10$, $\hat{\sigma}_{j|h}^2 = 8.86$, $\hat{\sigma}_{t|i}^2 = 0.273$ and $\hat{\sigma}_{t|h}^2 = 1.385$. This gives $a(\hat{\boldsymbol{\Sigma}}) \approx 0.28$, implying that $\hat{\boldsymbol{\beta}}^{sub}$ is approximately four times larger than $\hat{\boldsymbol{\beta}}$.

In our example the item correlation estimated by GEE is -0.26 (between items 1 and 2),

0.213 (1 and 3) and -0.16 (2 and 3) indicating a mix of negative and positive correlations. However, GLMM (10) assumes non-negative correlations (see Section 3). To impose positive correlations between all items, we applied a trick by transforming the 0/1 binary response to a 1/0 response for item 2. That is, positive responses become negative and negative responses become positive. Note this transformation changes the sign of the estimates for item 2. To make them comparable with the GEE method, the estimates were multiplied by -1. The trick works for data with these 3 items, but generally it might not work for $c \geq 2$. For example if the correlation between the items 2 and 3 would be positive instead, we could not apply this trick.

6 Discussion

This article focuses on the marginal and subject-specific approaches for modelling repeated multiple responses. For the marginal model approach, our main attention was directed towards quasi-likelihood methods, such as GEE and ALR, 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 are often infeasible due to a large value of 2^{cT} and the associated sparseness of the data. For the subject-specific approach, a GLMM takes the dependence among items and time points through the distribution of random effects into account. However it implies non-negative associations across different items due to the simple structure of the joint distribution. The simulation study showed this might result in not maintaining the type I error and therefore might lead to false statistical inference.

In general, the GEE method is widely available in all common statistical packages. Choosing a working correlation structure closer to the true situation can result in more efficient estimates. This paper recommends using the correlation models (6) to account for the two types of correlation, the item correlation and time-points correlation.

Standard GEE packages (eg. `geepack`) cannot fit correlation models as (6) directly. As an alternative, we proposed a 3-step method using existing GEE packages. The simulation study showed that this method works almost as well as estimating the correlation and mean model parameters jointly, and even performs well under other association models. The 3-step method is a trick that enables us to use existing software and avoids writing new code.

Common working correlations of the GEE method assume an equal correlation between responses Y_{ijt} and $Y_{ij't}$ for all subjects i. A possible extension is the group-wise method suggested by Suesse (2008), which assumes responses are equally correlated within the same group, but the strength of the correlation differs between groups. Grouping could naturally occur through variables such as gender. Modelling the correlation 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 approaches and leads often to more efficient estimates, provided that group-sizes are large and the number of parameters of the correlation model is relatively small. We 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 the choice of standard correlation structures, which assume equal correlations for all subjects.

Although both GEE and GLMM methods seem similar and both contain fixed effect parameters, namely $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^{sub}$, one does not imply the other. For our example, we are interested in how the probability of paying fees to doctors and extras (FD), paying private health insurance (PHI), and paying for holidays (HOL) depends on different factors. For instance, a factor of interest for a general population might be gender. Comparing females with males does not only refer to the household but to the general population. Therefore we are interested in a population-averaged effect and the marginal model is appropriate. The practitioner needs to be aware of the research question at hand to decide which approach should be applied. Generally speaking, the marginal models seem to be more useful than the subject–specific models in many applications. For discussions on the application of either

population-averaged or subject-specific models, see Neuhaus (1992) and Heagerty & Zeger (2000).

Other model approaches not considered here are marginalised GLMMs, transition models and log-linear models; for a good summary see Diggle et al. (2002). A marginalised GLMM has the advantage of a marginal interpretation, like GEE, and allows 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 , 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 the ML fitting become increasingly complex for large $c \cdot T$, as discussed in Section 2. Because most statistical packages do not offer to fit such models, it not as straightforward as for GEE or GLMM to apply such models.

Finally, we discuss the issue about missing data. The GEE method assumes data being missing completely at random (MCAR). Under the weaker assumption of data missing at random (MAR), GEE does not provide consistency in contrast to ML methods, such as GLMM. For our example, the standard GEE method seems reasonable, because a sub-case of MCAR allows missingness to depend on the observed covariates, e.g. time, age or sex. It is called the covariate-dependent missingness (Hedeker & Gibbons, 2006). If, however, missingness indeed depends on previous or current responses, then the general MAR case applies. GEE appproaches that try to account for MAR have been considered by Fitzmaurice et al. (1995), Ali & Talukder (2005) and Lipsitz et al. (2009).

Future research will investigate the impact of different functions on the correlations between responses across items and time points. Option 1 (6) suggests using a product, but other functions might be more appropriate. Also, an `R`-package might be developed for the

3-step method, generalising the method to more than two levels.

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. (2007). Modeling association between two or more categorical variables that allow for multiple category choices. *Communications in Statistics-Theory and Methods*, **36**(1-4), 433–451.

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.

By, K., Qaqish, B. F., & Preisser, J. S. (2011). *orth: Multivariate Logistic Regressions Using Orthogonalized Residuals.* R package version 1.5.1.

Carey, V., Zeger, S. L., & Diggle, P. (1993). Modeling multivariate binary data with alternating logistic regressions. *Biometrika*, **80**(3), 517–526.

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

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

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.

Guo, G. & Zhao, H. X. (2000). Multilevel modeling for binary data. *Annual Review of Sociology*, **26**, 441–462.

Hartzel, J., Agresti, A., & Caffo, B. (2001). Multinomial logit random effects models. *Stat. Model.*, **1**, 81–102.

Heagerty, P. J. & Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, **15**(1), 1–19.

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.

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

Jokinen, J. (2009). *drm: Regression and association models for repeated categorical data*. R package version 0.5-8.

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.

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.

Lipsitz, S., Fitzmaurice, G., Ibrahim, J., Sinha, D., M., P., & S., L. (2009). Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: An application to aids data. *Journal of the Royal Statistical Society Series A-Statistics in Society*, **172**(1), 3–20.

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. *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.

Molenberghs, G. & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer series in statistics. Springer.

Neuhaus, J. M. (1992). Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research*, **1**, 249–273.

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

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.

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.

Suesse, T. & Liu, I. (2012). Mantel-haenszel estimators of odds ratios for stratified dependent binomial data. *Computational Statistics & Data Analysis*, **56**(9), 2705–2717.

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), 339–348.

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.

Zink, R. C. (2003). *Correlated Binary Regression Using Orthogonalized Residuals*. PhD thesis.

Table 2: Simulation Results for model A 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	Correlation Model		
n	Method	Model I	Model II	Model III
30	unstr	1.165, 89.8, 88.5*	1.158, 89.9, 88.7*	1.151, 90.2, 88.8*
30	exch	1.020, 93.6, 93.4	1.014, 93.3, 93.4	1.014, 87.7, 93.5
30	ind	1.020, 90.8, 93.3	1.016, 90.2, 93.4	1.015, 88.8, 93.6
30	item	1.020, 90.8, 90.1	1.016, 90.1, 89.5	1.011, 88.7, 88.1
30	time	1.006, 95.0, 93.4	1.006, 95.0, 93.4	1.007, 95.1, 93.6
30	opt1-j	1.002, 95.4, 93.4*	1.001, 95.1, 93.5*	0.997, 95.0, 93.6*
30	opt1-s	1.006, 95.0, 93.3	1.006, 94.9, 93.4	1.003, 95.0, 93.6
30	opt2-j	1.013, 93.1, 93.4*	1.011, 93.7, 93.5*	1.026, 94.1, 93.5°
30	opt2-s	1.017, 94.9, 93.4*	1.017, 94.8, 93.4*	1.164, 93.8, 93.7*
30	alr1	1.026, 94.7, 93.1*	1.024, 94.5, 93.1*	1.017, 94.3, 93.0*
30	alr2	1.017, 94.2, 93.0°	1.010, 93.8, 93.1°	1.010, 92.7, 92.7•
100	unstr	1.056, 93.8, 93.5*	1.058, 93.7, 93.4	1.055, 93.6, 93.3*
100	exch	1.021, 93.4, 94.6	1.014, 93.2, 94.7	1.015, 87.2, 94.6
100	ind	1.020, 90.3, 94.6	1.013, 89.8, 94.6	1.016, 88.1, 94.6
100	item	1.020, 90.3, 90.1	1.013, 89.7, 89.5	1.013, 88.1, 87.9
100	time	1.004, 95.1, 94.6	1.003, 95.0, 94.7	1.009, 94.9, 94.6
100	opt1-j	1.003, 95.4, 94.6*	1.004, 95.0, 94.6*	1.003, 94.8, 94.6*
100	opt1-s	1.003, 95.1, 94.6	1.003, 95.0, 94.7	1.006, 94.9, 94.6
100	opt2-j	1.012, 93.4, 94.6*	1.003, 94.6, 94.6*	1.058, 94.2, 94.6*
100	opt2-s	1.006, 95.0, 94.5	1.003, 95.0, 94.6	1.159, 93.5, 94.6*
100	alr1	1.021, 94.9, 94.5	1.015, 94.7, 94.6	1.014, 94.6, 94.5*
100	alr2	1.020, 94.3, 94.5*	1.015, 93.9, 94.6*	1.049, 92.3, 94.4•
500	unstr	1.012, 94.9, 94.8	1.009, 94.8, 94.7	1.019, 94.8, 94.7
500	exch	1.019, 93.4, 94.9	1.010, 93.4, 95.2	1.014, 87.5, 94.9
500	ind	1.021, 90.5, 94.9	1.010, 89.9, 95.0	1.016, 88.2, 94.9
500	item	1.018, 90.5, 90.5	1.009, 89.8, 89.9	1.012, 88.2, 88.2
500	time	1.004, 95.0, 94.8	1.003, 95.2, 95.0	1.010, 95.0, 94.9
500	opt1-j	1.000, 95.3, 94.9*	1.002, 95.2, 95.1	1.006, 94.9, 95.0*
500	opt1-s	1.002, 95.0, 94.9	1.002, 95.2, 95.1	1.006, 95.0, 95.0
500	opt2-j	1.006, 94.3, 94.9*	1.001, 95.1, 95.1	1.053, 94.3, 95.0*
500	opt2-s	1.004, 95.1, 94.9	1.001, 95.0, 95.1	1.053, 94.4, 95.0
500	alr1	1.018, 95.0, 95.0	1.012, 95.0, 95.1	1.012, 94.8, 94.9
500	alr2	1.017, 94.4, 95.0	1.011, 94.0, 95.0	1.172, 90.8, 95.1•

Non-convergence rate: 0% (no symbol), $0 - 10\%$ (*), $10 - 50\%$ (°), $> 50\%$ (•)

Table 3: Simulation Results for model B 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	Correlation Model		
n	Method	Model I	Model II	Model III
30	unstr	1.435, 85.5, 87.7$^{\bullet}$	1.410, 85.7, 87.6$^{\bullet}$	1.341, 81.5, 88.0°
30	exch	1.025, 91.7, 92.2*	1.010, 91.8, 92.5*	1.012, 91.6, 93.0*
30	ind	1.025, 80.9, 92.2*	1.010, 80.0, 92.5*	1.014, 93.9, 93.0
30	item	1.037, 86.8, 88.5*	1.017, 86.1, 88.0*	0.999, 88.7, 91.3*
30	time	0.905, 88.6, 90.0*	0.863, 88.4, 89.8*	0.948, 97.5, 97.6*
30	opt1-j	0.883, 93.7, 92.8°	0.855, 93.3, 92.7°	0.946, 93.7, 93.0°
30	opt1-s	0.916, 93.0, 92.5*	0.875, 93.0, 92.6*	0.945, 94.0, 93.0*
30	opt2-j	0.919, 90.5, 92.5*	0.875, 90.7, 92.6°	0.975, 93.8, 93.1°
30	opt2-s	0.955, 92.1, 92.6*	0.907, 91.8, 92.8*	1.149, 93.7, 93.2*
30	alr1	0.941, 93.1, 91.6°	0.939, 92.6, 91.7°	0.923, 95.3, 93.0°
30	alr2	1.020, 92.3, 91.1°	0.991, 91.8, 91.3°	0.930, 95.7, 93.5$^{\bullet}$
100	unstr	1.300, 89.8, 91.9°	1.304, 90.3, 91.9°	1.097, 90.8, 92.7*
100	exch	1.031, 93.6, 93.9	1.010, 93.5, 93.8	1.013, 93.6, 94.2
100	ind	1.031, 82.3, 93.9	1.010, 81.3, 93.8	1.013, 94.1, 94.2
100	item	1.033, 88.6, 89.4*	1.016, 88.0, 88.6*	1.012, 90.8, 91.6
100	time	1.013, 89.2, 89.8*	1.005, 88.7, 89.3*	1.015, 97.1, 97.4
100	opt1-j	1.011, 94.3, 94.0*	1.011, 93.7, 93.7*	1.015, 94.5, 94.2*
100	opt1-s	1.016, 94.0, 93.9*	1.011, 93.7, 93.8*	1.014, 94.9, 94.2
100	opt2-j	1.032, 91.3, 93.7*	1.013, 91.9, 93.8*	1.024, 95.3, 94.2*
100	opt2-s	1.076, 92.5, 93.8*	1.079, 92.3, 93.9*	1.040, 95.3, 94.2*
100	alr1	1.061, 91.9, 93.4*	1.044, 91.5, 93.4*	1.020, 95.1, 93.7*
100	alr2	1.049, 91.4, 93.3°	1.027, 90.5, 93.2°	1.041, 94.9, 95.3$^{\bullet}$
500	unstr	1.143, 94.1, 94.8*	1.154, 93.8, 94.6*	1.012, 94.4, 94.8*
500	exch	1.023, 95.5, 95.6	1.006, 95.4, 95.6	1.009, 94.6, 95.0
500	ind	1.023, 84.3, 95.6	1.006, 84.2, 95.6	1.009, 94.7, 95.0
500	item	1.021, 90.6, 91.0	1.010, 90.4, 90.4	1.009, 91.7, 92.0
500	time	1.010, 91.1, 91.2	1.006, 91.4, 91.3	1.005, 97.6, 97.9
500	opt1-j	1.007, 95.7, 95.5*	1.008, 95.2, 95.5*	1.004, 95.4, 94.9*
500	opt1-s	1.008, 95.3, 95.5	1.010, 95.2, 95.5	1.005, 95.7, 94.9
500	opt2-j	1.014, 93.9, 95.5*	1.003, 94.2, 95.6*	1.009, 96.2, 94.9*
500	opt2-s	1.014, 94.6, 95.5	1.008, 94.1, 95.6	1.010, 96.2, 94.9
500	alr1	1.064, 93.0, 95.5	1.055, 92.8, 95.4	1.005, 95.9, 94.8
500	alr2	1.042, 92.6, 95.3°	1.042, 92.0, 95.1°	1.178, 92.4, 95.4$^{\bullet}$

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

Table 4: Type I errors for β_{01} and β_{11} in model A for $n = 30, 100, 500$ based on naive variance followed by robust variance

	Fitting	Correlation Model		
n	Method	Model I	Model II	Model III
30	unstr	10.7, 10.6, 11.6, 12.9*	10.2, 10.4, 11.0, 12.2*	11.4, 10.6, 11.9, 12.6*
30	exch	6.59, 6.67, 6.13, 7.32	6.66, 6.77, 6.15, 7.09	12.9, 11.9, 7.07, 7.25
30	ind	9.20, 9.60, 6.13, 7.35	10.1, 9.76, 6.13, 7.01	12.1, 10.9, 7.01, 7.25
30	item	9.28, 9.72, 9.81, 10.2	10.1, 9.81, 10.6, 10.3	12.1, 11.1, 12.6, 11.7
30	time	5.04, 5.12, 6.03, 7.40	5.06, 5.28, 6.07, 7.00	6.77, 5.77, 6.90, 7.39
30	opt1-j	4.68, 4.77, 6.05, 7.35*	5.03, 5.28, 6.22, 6.88*	6.66, 5.83, 6.96, 7.44*
30	opt1-s	5.05, 5.17, 6.07, 7.40	5.12, 5.28, 6.17, 6.91	6.85, 5.88, 6.96, 7.34
30	opt2-j	6.92, 7.29, 6.05, 7.14*	6.33, 6.42, 5.97, 6.96*	7.39, 6.80, 7.09, 7.58°
30	opt2-s	5.13, 5.22, 6.04, 7.29*	5.05, 5.32, 6.04, 6.96*	6.85, 6.09, 7.20, 7.71*
30	alr1	4.82, 5.10, 6.05, 7.51*	4.92, 5.37, 6.31, 6.99*	6.88, 5.93, 7.01, 7.58*
30	alr2	5.38, 5.50, 6.05, 7.33°	5.73, 5.82, 6.03, 6.95°	8.14, 7.01, 6.74, 7.71•
30	glmm	5.26, 6.89*, −−, −−	5.44, 6.94*,−−, −−	8.87, 7.74*, −−, −−
100	unstr	6.13, 6.33, 6.13, 6.82*	5.90, 6.66, 5.90, 7.21	6.65, 8.31, 6.76, 8.25*
100	exch	6.26, 6.86, 4.83, 5.84	6.44, 7.13, 4.90, 5.89	13.4, 13.7, 5.43, 6.32
100	ind	9.75, 9.75, 4.86, 5.80	10.0, 10.5, 4.89, 5.81	12.6, 13.0, 5.32, 6.32
100	item	9.67, 9.71, 9.80, 9.87	10.1, 10.6, 10.2, 10.7	12.9, 13.5, 13.1, 13.7
100	time	4.70, 5.11, 4.99, 5.78	4.82, 5.30, 4.93, 5.70	5.98, 6.65, 5.48, 6.65
100	opt1-j	4.39, 4.91, 4.89, 5.88*	4.75, 5.35, 4.91, 5.75*	6.20, 6.93, 5.37, 6.43*
100	opt1-s	4.72, 5.17, 4.87, 5.92	4.80, 5.35, 4.92, 5.70	6.04, 6.93, 5.32, 6.48
100	opt2-j	6.28, 6.72, 4.89, 5.80*	5.25, 5.66, 4.92, 5.74*	6.09, 6.87, 5.43, 6.37*
100	opt2-s	4.69, 5.23, 4.95, 5.87	4.82, 5.26, 4.91, 5.77	6.04, 6.81, 5.43, 6.37*
100	alr1	4.17, 4.80, 4.82, 5.77	4.49, 5.10, 4.96, 5.64	5.71, 6.20, 5.71, 5.98*
100	alr2	5.07, 5.55, 5.01, 5.81*	5.33, 5.99, 5.03, 5.77*	8.42, 8.86, 4.93, 5.98•
100	glmm	4.90, 6.70, −−, −−	5.01, 6.91, −−, −−	7.70, 8.48*, −−, −−
500	unstr	5.68, 4.62, 5.73, 4.87	6.35, 5.10, 6.35, 5.25	5.76, 4.97, 5.68, 5.09
500	exch	7.14, 6.13, 5.68, 4.92	7.65, 6.30, 5.75, 4.90	14.0, 12.6, 5.51, 4.88
500	ind	10.3, 9.10, 5.68, 4.72	11.3, 9.70, 5.95, 4.80	13.1, 12.0, 5.55, 4.88
500	item	10.3, 9.15, 10.4, 9.20	11.3, 9.65, 11.3, 9.70	13.0, 12.0, 13.1, 11.9
500	time	5.28, 4.67, 5.38, 4.92	5.80, 4.70, 5.90, 5.00	6.09, 5.01, 5.68, 4.88
500	opt1-j	5.18, 4.42, 5.33, 4.87*	5.80, 4.70, 5.80, 5.00	6.18, 5.22, 5.68, 4.63*
500	opt1-s	5.28, 4.62, 5.28, 4.92	5.85, 4.65, 5.80, 5.05	6.14, 5.09, 5.68, 4.63
500	opt2-j	6.08, 5.38, 5.48, 4.92*	5.80, 4.65, 5.08, 4.90	6.68, 5.80, 5.93, 4.72*
500	opt2-s	5.33, 4.77, 5.38, 4.97	5.80, 4.65, 5.80, 4.90	6.72, 5.76, 5.88, 4.72
500	alr1	5.28, 4.27, 5.53, 4.72	5.55, 4.35, 6.00, 4.70	6.18, 4.80, 5.84, 4.88
500	alr2	5.68, 4.92, 5.43, 4.97	6.35, 5.60, 5.90, 5.15	11.4, 9.47, 5.51, 4.47•
500	glmm	5.68, 5.93, −−, −−	6.35, 6.05, −−, −−	7.89, 6.84, −−, −−

Non-convergence rate: 0% (no symbol), $0-10\%$ (*), $10-50\%$ (°), $>50\%$ (•)

Table 5: Type I errors for β_{01} and β_{11} in model A* for $n = 30, 100, 500$ based on naive variance followed by robust variance

Fitting Method	Cluster Size n		
	30	100	500
unstr	9.76, 9.35, 10.7, 12.1*	6.05, 6.38, 6.15, 6.61	5.58, 5.70, 5.58, 5.72
exch	7.80, 7.30, 6.01, 7.05	7.85, 8.48, 4.88, 5.68	8.91, 9.15, 5.40, 5.44
ind	9.28, 8.89, 5.98, 7.02	9.39, 10.2, 4.89, 5.71	10.5, 10.9, 5.40, 5.40
item	9.28, 8.86, 9.63, 9.56	9.39, 10.2, 9.52, 10.2	10.5, 10.9, 10.5, 11.0
time	5.32, 5.30, 5.92, 7.00	5.26, 5.58, 4.95, 5.61	5.92, 5.88, 5.42, 5.40
opt1-j	6.03, 5.86, 5.90, 6.99*	5.86, 6.17, 4.99, 5.57*	6.70, 6.62, 5.52, 5.32*
opt1-s	5.28, 5.30, 5.91, 6.99	5.27, 5.61, 4.99, 5.58	5.92, 5.90, 5.44, 5.42
opt2-j	5.48, 5.36, 5.94, 7.00*	5.19, 5.54, 4.98, 5.59*	5.94, 5.90, 5.40, 5.40
opt2-s	5.30, 5.28, 5.94, 7.03	5.19, 5.54, 4.98, 5.59	5.94, 5.90, 5.40, 5.40
alr1	5.66, 5.49, 6.13, 7.08	5.14, 5.54, 4.97, 5.61	5.84, 5.80, 5.60, 5.38
alr2	6.81, 6.67, 5.90, 7.02*	6.54, 6.74, 4.96, 5.63	7.08, 7.08, 5.32, 5.42
glmm	6.38, 6.84*, $--$, $--$	6.18, 6.29, $--$, $--$	5.00, 5.63, $--$, $--$

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

Table 6: Results for model B and $n = 30, 100, 500$ - 100 $\times$MSE and coverage of 95% confidence interval based on naive variance followed by robust variance for the parameters β_0 and β_1

	Fitting	Correlation Model	
n	Method	Model I	Model III
30	unstr	11.93, 47.778, 85.3, 86.6, 88.9, 88.7$^\bullet$	6.334, 21.652, 80.9, 79.0, 88.0, 87.6°
30	exch	8.646, 34.025, 93.6, 92.6, 94.4, 92.2*	4.462, 16.654, 93.5, 93.1, 93.3, 93.8*
30	ind	8.646, 34.025, 84.9, 84.5, 94.4, 92.2*	4.462, 16.701, 94.4, 94.4, 93.3, 93.8
30	item	8.805, 34.344, 90.2, 89.8, 91.7, 90.6*	4.458, 16.375, 92.1, 91.8, 91.7, 93.1*
30	time	7.980, 29.668, 90.8, 90.1, 92.0, 90.2*	4.292, 15.494, 96.8, 96.5, 97.2, 96.1*
30	opt1-j	7.939, 28.829, 94.9, 94.0, 94.3, 92.0°	4.334, 15.403, 94.6, 94.5, 93.1, 94.0°
30	opt1-s	8.122, 29.992, 94.5, 93.8, 94.2, 92.1*	4.305, 15.401, 94.8, 94.5, 93.1, 93.9*
30	opt2-j	8.144, 30.119, 91.8, 91.2, 94.4, 92.3*	4.440, 15.907, 94.9, 94.5, 93.1, 93.9°
30	opt2-s	8.423, 31.339, 94.1, 93.1, 94.3, 92.0*	5.222, 18.749, 95.3, 95.0, 93.1, 94.0*
30	alr1	8.421, 30.763, 94.9, 92.7, 93.7, 91.7°	4.301, 14.963, 94.4, 95.2, 92.3, 93.1°
30	alr2	8.550, 33.904, 94.5, 92.9, 94.3, 92.2°	4.608, 14.804, 94.9, 95.7, 92.6, 93.5$^\bullet$
100	unstr	3.372, 11.01, 88.8, 89.5, 91.6, 90.7°	1.318, 4.950, 93.4, 91.0, 94.3, 93.2*
100	exch	2.715, 8.696, 93.5, 93.0, 93.9, 93.0	1.210, 4.576, 95.2, 93.6, 95.8, 94.4
100	ind	2.715, 8.696, 83.4, 82.9, 93.9, 93.0	1.210, 4.576, 95.2, 93.6, 95.8, 94.4
100	item	2.720, 8.716, 89.1, 88.6, 89.9, 88.8*	1.209, 4.573, 92.8, 91.2, 93.4, 92.5
100	time	2.679, 8.537, 89.4, 89.0, 90.3, 89.2*	1.208, 4.588, 97.3, 96.8, 98.1, 97.0
100	opt1-j	2.670, 8.522, 94.3, 93.9, 94.0, 93.2*	1.208, 4.592, 95.7, 94.6, 95.9, 94.7*
100	opt1-s	2.685, 8.563, 94.0, 93.6, 93.9, 93.1*	1.208, 4.584, 96.0, 94.9, 95.9, 94.8
100	opt2-j	2.717, 8.710, 91.6, 90.9, 94.0, 92.9*	1.221, 4.633, 96.2, 94.9, 96.0, 94.5*
100	opt2-s	2.869, 9.042, 93.3, 92.4, 93.9, 93.0*	1.232, 4.708, 96.4, 95.4, 95.9, 94.7*
100	alr1	2.773, 8.971, 93.1, 90.5, 93.7, 92.5*	1.223, 4.605, 96.4, 95.4, 95.3, 94.6*
100	alr2	2.680, 8.928, 92.7, 90.1, 93.9, 92.7°	1.325, 4.622, 94.7, 95.1, 96.0, 94.6$^\bullet$
500	unstr	0.582, 1.640, 93.7, 94.4, 94.5, 94.9*	0.243, 0.887, 95.2, 93.8, 94.9, 94.9*
500	exch	0.530, 1.459, 94.7, 95.8, 95.0, 95.7	0.242, 0.883, 95.7, 93.7, 95.4, 94.6
500	ind	0.530, 1.459, 83.1, 85.3, 95.0, 95.7	0.242, 0.883, 95.6, 93.7, 95.4, 94.6
500	item	0.528, 1.457, 90.2, 91.4, 90.3, 91.9	0.242, 0.883, 92.9, 91.1, 92.1, 92.2
500	time	0.524, 1.440, 90.5, 91.7, 91.3, 91.5	0.242, 0.879, 97.9, 97.2, 98.3, 97.2
500	opt1-j	0.520, 1.438, 95.1, 95.7, 95.1, 95.5*	0.241, 0.879, 96.1, 94.6, 95.5, 94.6*
500	opt1-s	0.522, 1.439, 94.8, 95.6, 95.1, 95.5	0.242, 0.879, 96.2, 94.9, 95.5, 94.6
500	opt2-j	0.525, 1.447, 93.3, 94.2, 95.3, 95.3*	0.243, 0.882, 96.7, 95.6, 95.4, 94.8*
500	opt2-s	0.527, 1.445, 94.0, 95.1, 95.3, 95.5	0.243, 0.883, 96.7, 95.5, 95.5, 94.8
500	alr1	0.539, 1.530, 93.9, 92.4, 95.4, 95.4	0.240, 0.881, 96.6, 95.3, 95.4, 94.3
500	alr2	0.535, 1.491, 93.3, 91.9, 95.4, 95.4°	0.361, 0.952, 90.3, 94.7, 95.8, 95.1$^\bullet$

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

Table 7: Results for fees paid to doctors and other health care professionals ('Extras')

	GEE			ALR			GLMM		
Variable	estimate	(s.e.)	p-value	estimate	(s.e.)	p-value	estimate	(s.e.)	p-value
Intercept	1.605	(0.194)	$1e-16$	1.351	(0.212)	$2e-10$	-4.447	(0.463)	0.0000
Age - baseline: 18-24									
25-34	0.200	(0.075)	0.008	0.212	(0.078)	0.006	-0.525	(0.169)	$3e-04$
35-49	0.248	(0.065)	$1e-04$	0.263	(0.069)	$1e-04$	-0.664	(0.155)	$2e-07$
50-74	0.485	(0.064)	$5e-14$	0.485	(0.069)	$3e-12$	-1.081	(0.158)	$1e-15$
>74	0.122	(0.093)	0.188	0.034	(0.103)	0.743	0.029	(0.229)	0.872
Alcohol drinking status - baseline: abstainer									
ex-drinker	0.222	(0.087)	0.011	0.295	(0.095)	0.002	-0.771	(0.200)	$7e-04$
low risk	0.485	(0.067)	$4e-13$	0.556	(0.073)	$3e-14$	-1.149	(0.161)	$1e-10$
risky	0.521	(0.090)	$8e-09$	0.623	(0.099)	$3e-10$	-1.200	(0.216)	$4e-08$
high risk	0.372	(0.116)	0.001	0.340	(0.128)	0.008	-0.600	(0.277)	0.004
Education - baseline: higher education									
some education	-0.565	(0.055)	0.0000	-0.668	(0.058)	0.0000	1.102	(0.145)	0.0000
no education	-0.922	(0.059)	0.0000	-1.076	(0.063)	0.0000	2.019	(0.156)	0.0000
Baseline: English first language learned									
English not first language	-0.005	(0.048)	0.923	-0.002	(0.051)	0.963	-0.069	(0.112)	0.246
Gross weekly income	$2e-04$	$(4e-05)$	$4e-08$	$3e-04$	$(4e-05)$	$6e-12$	$-5e-04$	$(1e-04)$	$2e-12$
Labour Force Status - baseline: employed									
unemployed	-0.281	(0.115)	0.015	-0.037	(0.136)	0.787	0.519	(0.228)	0.031
Not in the labour force	-0.313	(0.055)	$1e-08$	-0.237	(0.062)	$1e-04$	0.737	(0.140)	$7e-10$
Satisfaction Scores 0-100									
Satisfaction - Life	-0.035	(0.012)	0.003	-0.036	(0.012)	0.003	0.062	(0.026)	$5e-04$
Satisfaction - Partner	-0.032	(0.015)	0.032	-0.020	(0.016)	0.214	0.064	(0.031)	$3e-06$
Marital Status - baseline: married									
De facto	-0.660	(0.067)	0.0000	-0.684	(0.070)	0.0000	1.658	(0.172)	0.0000
Separated	-1.002	(0.105)	0.0000	-0.911	(0.118)	$1e-14$	2.194	(0.266)	0.0000
Divorced	-0.936	(0.077)	0.0000	-0.875	(0.082)	0.0000	2.227	(0.207)	0.0000
Widowed	-0.925	(0.085)	0.0000	-0.798	(0.093)	0.0000	2.291	(0.235)	0.0000
Never married and not de facto	-1.128	(0.066)	0.0000	-1.100	(0.067)	0.0000	2.784	(0.172)	0.0000
Long Term Health Conditions (LTHC)									
Pain	-0.139	(0.068)	0.040	-0.216	(0.074)	0.003	0.433	(0.160)	0.005
Shortness of Breath	-0.033	(0.078)	0.668	0.024	(0.083)	0.771	0.111	(0.190)	0.611
Female	0.398	(0.039)	0.0000	0.444	(0.041)	0.0000	-1.044	(0.099)	0.0000
Smoking Status - baseline: do not smoke									
no longer smoke	-0.035	(0.047)	0.454	-0.078	(0.051)	0.131	0.087	(0.119)	0.410
smoke daily	-0.451	(0.072)	$4e-10$	-0.607	(0.078)	$7e-15$	1.032	(0.171)	$4e-07$
smoke weekly	-0.191	(0.131)	0.143	-0.395	(0.138)	0.004	0.595	(0.259)	0.023
less often than weekly	-0.239	(0.131)	0.068	-0.433	(0.133)	0.001	0.745	(0.322)	0.091
Race - baseline: Not indigenous									
Indigenous	-0.544	(0.132)	$4e-05$	-0.691	(0.132)	$2e-07$	1.546	(0.361)	$4e-05$
Cigarettes	$-5e-04$	$(5e-04)$	0.326	$-5e-04$	$(6e-04)$	0.433	0.002	(0.001)	0.513
Health Scores 0-100									
Physical functioning	0.004	(0.001)	$1e-04$	0.005	(0.001)	$4e-05$	-0.012	(0.002)	$2e-07$
Bodily Pain	-0.005	(0.001)	$1e-06$	-0.006	(0.001)	$9e-08$	0.015	(0.002)	$5e-09$
Vitality	-0.003	(0.001)	0.033	-0.003	(0.002)	0.081	0.009	(0.003)	$3e-09$
Mental Health	0.006	(0.002)	$3e-04$	0.006	(0.002)	$2e-04$	-0.014	(0.003)	$7e-15$

Table 8: Results for Private Health Insurance

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	−0.753	(0.112)	$2e-11$	−0.441	(0.126)	$5e-04$	−1.386	(0.408)	0.079
Age - baseline: 18-24									
25-34	−0.046	(0.052)	0.376	−0.070	(0.052)	0.182	−0.414	(0.156)	0.196
35-49	−0.076	(0.037)	0.039	−0.224	(0.043)	$3e-07$	−0.859	(0.143)	$4e-07$
50-74	−0.246	(0.038)	$6e-11$	−0.701	(0.046)	0.0000	−2.018	(0.148)	0.0000
>74	−0.044	(0.056)	0.429	−0.342	(0.072)	$2e-06$	−0.787	(0.233)	$6e-05$
Alcohol drinking status - baseline: abstainer									
ex-drinker	−0.010	(0.048)	0.834	−0.012	(0.054)	0.826	0.077	(0.200)	0.532
low risk	−0.124	(0.042)	0.003	−0.266	(0.047)	$2e-08$	−0.620	(0.155)	0.030
risky	−0.161	(0.053)	0.002	−0.321	(0.060)	$7e-08$	−0.742	(0.204)	0.007
high risk	−0.202	(0.067)	0.003	−0.317	(0.079)	$6e-05$	−0.606	(0.275)	0.211
Dependent Person	0.019	(0.111)	0.862	−0.178	(0.105)	0.091	−0.503	(0.282)	0.025
Education - baseline: higher education									
some education	0.682	(0.042)	0.0000	0.850	(0.048)	0.0000	1.998	(0.119)	0.0000
no education	0.881	(0.046)	0.0000	1.053	(0.054)	0.0000	2.576	(0.135)	0.0000
Baseline: English first language learned									
English not first language	0.034	(0.022)	0.124	0.162	(0.025)	$2e-10$	0.483	(0.101)	$2e-04$
Gross weekly income	$-1e-04$	$(2e-05)$	$3e-09$	$-4e-04$	$(3e-05)$	0.0000	−0.001	$(9e-05)$	0.0000
Labour Force Status - baseline: employed									
unemployed	0.009	(0.059)	0.879	0.067	(0.069)	0.332	0.232	(0.262)	0.538
Not in the labour force	0.017	(0.030)	0.569	$-3e-04$	(0.037)	0.994	0.010	(0.127)	0.471
Satisfaction Scores 0-100									
Satisfaction - Partner	0.012	(0.006)	0.033	0.020	(0.006)	0.002	0.057	(0.023)	0.437
Mental health	0.045	(0.068)	0.508	0.155	(0.088)	0.078	0.438	(0.349)	0.352
Marital Status - baseline: married									
De facto	0.495	(0.052)	0.0000	0.661	(0.054)	0.0000	2.223	(0.153)	0.0000
Separated	0.530	(0.074)	$1e-12$	0.775	(0.077)	0.0000	2.524	(0.241)	0.0000
Divorced	0.678	(0.059)	0.0000	0.939	(0.067)	0.0000	3.076	(0.194)	0.0000
Widowed	0.487	(0.061)	$1e-15$	0.578	(0.078)	$1e-13$	1.978	(0.225)	$8e-15$
Never married and not de facto	0.885	(0.060)	0.0000	0.982	(0.059)	0.0000	3.219	(0.168)	0.0000
Long Term Health Conditions (LTHC)									
Pain	−0.017	(0.029)	0.554	0.020	(0.034)	0.556	0.182	(0.157)	0.364
Shortness of Breath	0.007	(0.041)	0.871	0.084	(0.046)	0.066	0.383	(0.199)	0.049
Female	−0.055	(0.024)	0.022	−0.131	(0.031)	$3e-05$	−0.316	(0.088)	$4e-07$
Smoking Status - baseline: do not smoke									
no longer smoke	0.058	(0.031)	0.063	0.153	(0.034)	$9e-06$	0.512	(0.104)	0.001
smoke daily	0.364	(0.043)	0.0000	0.774	(0.048)	0.0000	2.036	(0.138)	0.0000
smoke weekly	0.256	(0.065)	$8e-05$	0.525	(0.072)	$3e-13$	1.421	(0.258)	$2e-04$
less often than weekly	0.103	(0.070)	0.143	0.258	(0.077)	$9e-04$	0.771	(0.293)	0.056
Race - baseline: Not indigenous									
Indigenous	0.650	(0.119)	$4e-08$	0.745	(0.157)	$2e-06$	1.312	(0.440)	0.004
Wave	−0.013	(0.006)	0.035	0.006	(0.007)	0.375	−0.018	(0.027)	0.433
Developed LTHC at T-1	−0.030	(0.042)	0.485	−0.040	(0.042)	0.345	−0.195	(0.228)	0.468
Health Scores 0-100									
Physical functioning	−0.001	$(5e-04)$	0.019	−0.003	$(6e-04)$	$1e-05$	−0.007	(0.002)	0.364
Mental Health	$-8e-04$	$(6e-04)$	0.197	−0.003	$(7e-04)$	$5e-05$	−0.008	(0.003)	0.004

Table 9: Results for holidays and holiday travel costs

Variable	GEE estimate	GEE (s.e.)	GEE p-value	ALR estimate	ALR (s.e.)	ALR p-value	GLMM estimate	GLMM (s.e.)	GLMM p-value
Intercept	−0.357	(0.152)	0.019	−0.324	(0.177)	0.067	0.638	(0.378)	0.003
Age - baseline: 18-24									
25-34	0.220	(0.056)	$7e-05$	0.212	(0.065)	0.001	−0.659	(0.136)	$2e-07$
35-49	0.164	(0.048)	$7e-04$	0.143	(0.056)	0.011	−0.414	(0.124)	0.044
50-74	0.260	(0.048)	$6e-08$	0.284	(0.057)	$6e-07$	−0.515	(0.124)	0.001
>74	−0.253	(0.074)	$7e-04$	−0.258	(0.090)	0.004	1.005	(0.192)	$2e-06$
Alcohol drinking status - baseline: abstainer									
ex-drinker	0.207	(0.065)	0.002	0.302	(0.081)	$2e-04$	−0.376	(0.164)	0.046
low risk	0.589	(0.052)	0.0000	0.700	(0.063)	0.0000	−1.461	(0.131)	0.0000
risky	0.685	(0.069)	0.0000	0.771	(0.084)	0.0000	−1.710	(0.174)	0.0000
high risk	0.655	(0.090)	$3e-13$	0.707	(0.111)	$2e-10$	−1.533	(0.229)	$2e-09$
Dependent Person	−0.016	(0.098)	0.874	−0.144	(0.111)	0.196	0.146	(0.246)	0.053
Developed LTHC	0.092	(0.049)	0.060	0.087	(0.060)	0.145	−0.331	(0.123)	0.008
Education - baseline: higher education									
some education	−0.556	(0.041)	0.0000	−0.696	(0.048)	0.0000	1.319	(0.114)	0.0000
no education	−0.850	(0.044)	0.0000	−1.039	(0.052)	0.0000	2.123	(0.125)	0.0000
Baseline: English first language learned									
English not first language	−0.154	(0.035)	$1e-05$	−0.200	(0.041)	$1e-06$	0.490	(0.086)	$8e-06$
Gross weekly income	$3e-04$	$(3e-05)$	0.0000	$3e-04$	$(4e-05)$	0.0000	$-6e-04$	$(8e-05)$	0.0000
Labour Force Status - baseline: employed									
unemployed	−0.211	(0.084)	0.012	−0.229	(0.097)	0.018	0.703	(0.194)	$4e-04$
Not in the labour force	−0.146	(0.041)	$4e-04$	−0.150	(0.049)	0.002	0.487	(0.106)	0.056
Satisfaction Scores 0-100									
Satisfaction - Partner	0.054	(0.008)	$3e-11$	0.055	(0.010)	$1e-08$	−0.169	(0.020)	$2e-14$
Satisfaction - Life	0.014	(0.010)	0.180	0.018	(0.013)	0.148	−0.039	(0.025)	0.704
Mental health	−0.178	(0.113)	0.116	−0.223	(0.126)	0.077	0.720	(0.269)	0.037
Marital Status - baseline: married									
De facto	−0.417	(0.054)	$1e-14$	−0.481	(0.060)	$7e-16$	1.083	(0.139)	0.0000
Separated	−0.629	(0.084)	$9e-14$	−0.724	(0.099)	$3e-13$	1.622	(0.201)	$2e-10$
Divorced	−0.793	(0.064)	0.0000	−0.847	(0.073)	0.0000	2.187	(0.167)	0.0000
Widowed	−0.511	(0.075)	$9e-12$	−0.509	(0.083)	$9e-10$	1.542	(0.202)	$5e-13$
Never married and not de facto	−0.817	(0.055)	0.0000	−0.859	(0.061)	0.0000	2.194	(0.148)	0.0000
Long Term Health Conditions (LTHC)									
Pain	−0.080	(0.047)	0.088	−0.085	(0.060)	0.155	0.214	(0.125)	0.124
Shortness of Breath	−0.200	(0.060)	$9e-04$	−0.217	(0.076)	0.004	0.448	(0.158)	0.023
Female	0.101	(0.028)	$3e-04$	0.103	(0.033)	0.002	−0.315	(0.080)	0.006
Smoking Status - baseline: do not smoke									
no longer smoke	−0.080	(0.035)	0.024	−0.104	(0.046)	0.024	0.127	(0.093)	0.293
smoke daily	−0.356	(0.056)	$2e-10$	−0.449	(0.065)	$4e-12$	0.775	(0.145)	$1e-08$
smoke weekly	−0.122	(0.090)	0.174	−0.221	(0.105)	0.035	0.189	(0.211)	0.785
less often than weekly	−0.027	(0.098)	0.780	−0.063	(0.117)	0.590	0.015	(0.246)	0.993
Race - baseline: Not indigenous									
Indigenous	−0.405	(0.127)	0.001	−0.483	(0.135)	$4e-04$	1.103	(0.319)	0.010
Number Cigarettes	−0.001	$(4e-04)$	0.002	−0.001	$(5e-04)$	0.007	0.004	(0.001)	0.127
Wave	−0.035	(0.010)	$6e-04$	−0.052	(0.011)	$3e-06$	0.101	(0.022)	0.085
Health Scores 0-100									
Physical functioning	0.004	$(7e-04)$	$7e-07$	0.004	$(9e-04)$	$1e-05$	−0.012	(0.002)	$8e-12$
Vitality	0.003	(0.001)	0.005	0.003	(0.001)	0.006	−0.006	(0.002)	0.012
Mental Health	0.004	(0.001)	$7e-05$	0.004	(0.001)	0.001	−0.010	(0.003)	$6e-04$