A COMPARISON OF STATISTICAL METHODS FOR THE ANALYSIS OF BINARY REPEATED MEASURES DATA WITH ADDITIONAL HIERARCHICAL STRUCTURE

ELMABROK MASAOD^{*}

*School of Epidemiology and Public Health, University of Ottawa
451 Smyth Road, Ottawa, Ontario, K1H 8M5, Canada
Email: emasaoud@uottawa.ca*

HENRIK STRYHN

*Department of Health Management, Atlantic Veterinary College
University of Prince Edward Island, Charlottetown, PEI C1A 4P3, Canada
Email: hstryhn@upe.ca*

SUMMARY

The objective of the study was to compare statistical methods for the analysis of binary repeated measures data with an additional hierarchical level. Random effects true models with autocorrelated ($\rho = 1, 0.9$ or 0.5) subject random effects were used in this simulation study. The settings of the simulation were chosen to reflect a real veterinary somatic cell count dataset, except that the within-subject time series were balanced, complete and of fixed length (4 or 8 time points). Four fixed effects parameters were studied: binary predictors at the subject and cluster levels, respectively, a linear time effect, and the intercept. The following marginal and random effects statistical procedures were considered: ordinary logistic regression (OLR), alternating logistic regression (ALR), generalized estimating equations (GEE), marginal quasi-likelihood (MQL), penalized quasi-likelihood (PQL), pseudo likelihood (REPL), maximum likelihood (ML) estimation and Bayesian Markov chain Monte Carlo (MCMC). The performance of these estimation procedures was compared specifically for the four fixed parameters as well as variance and correlation parameters. The findings of this study indicate that in data generated by random intercept models ($\rho = 1$), the ML and MCMC procedures performed well and had fairly similar estimation errors. The PQL regression estimates were attenuated while the variance estimates were less accurate than ML and MCMC, but the direction of the bias depended on whether binomial or extra-binomial dispersion was assumed. In datasets with autocorrelation ($\rho < 1$), random effects estimates procedures gave downwards biased estimates, while marginal estimates were little affected by the presence of autocorrelation. The results also indicate that in addition to ALR, a GEE procedure that accounts for clustering at the highest hierarchical level is sufficient.

Keywords and phrases: Hierarchical structure; Marginal models; Random effects models; Binary repeated measures; Simulation.

^{*} Corresponding author

© Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh.

1 Introduction

Binary repeated measures data with additional hierarchical structure are data with multiple records over time on the same subjects (e.g., animals or farms), which in addition are nested within some (physical) clusters (e.g., hospitals, herds, provinces). In multi-level modelling terminology (Snijders and Bosker, 2012) this may be termed three-level repeated measures data, with observations corresponding to level one and clusters to level three. Such data structures are encountered across a wide range of applications in veterinary and human epidemiology. Our motivating example was a literature dataset of Dohoo *et al.* (2009, Chapter 31) on records of presence or absence of bacteria in monthly milk samples from cows housed in multiple herds. Thus, the hierarchical structure is the clustering of cows in herds, and the repeated measures are the monthly test records based on the milk samples. Data with this structure are common in studies of dairy cow udder health (Green *et al.*, 2004) and (Olde Riekerink *et al.*, 2007). Some examples from human preventive medicine include the effects of air pollution on school absences in the southern California Children's Health study (Virginie *et al.*, 2005), and the sickness episodes for workers over time (Preisner *et al.*, 2003).

Binary records made on the same subjects, nested within clusters, over time are likely to be correlated (Schukken *et al.*, 2003) or clustered (Dohoo *et al.*, 2009, Chapters 20–23). A within-subject dependence violates the basic assumption of logistic regression that observations are independent, and may, if not accounted for, lead to biases in parameter estimates and standard errors (Diggle *et al.*, 2002, Chapter 7) and (Dohoo *et al.*, 2009, Chapter 20). Such data structures challenge the statistical methods to hold its properties, such as asymptotic unbiasedness and nominal confidence interval coverage.

Numerous procedures (models) have been proposed for the analysis of binary repeated measures data; a basic distinction is between marginal (population-averaged, or PA) and random effects (subject-specific, or SS) models ((Diggle *et al.*, 2002, Chapters 7–10) and (Neuhaus, 1992)). A large body of literature on statistical methods of binary repeated measures data have discussed the choice between these model types and specific procedures, see for example Diggle *et al.* (2002, Chapters: 7–11) or a simulation study by Masaoud and Stryhn (2010). However, the added hierarchical structure poses problems for procedures of both types. Ignoring the top or intermediate hierarchical structure in multilevel data analysis has an effect on the estimated variance components and that standard errors of regression coefficients estimators may be overestimated (Moerbeek, 2004). Moreover, the impact of the added hierarchical structure would intuitively be expected to differ not only between estimation approaches but also between types of parameters in a model. The fixed part of a model could contain predictors at all three levels: the cluster level, the subject level, and observation (within-subject) level. The random part of a model would involve variances and covariances.

In order to realistically reflect the choice an applied researcher faces when it comes to data analysis, only estimation procedures implemented in broadly accessible statistical software were considered for the study. Specifically, the following procedures were included: ordinary logistic regression (OLR), alternating logistic regression (ALR), generalized estimating equations (GEE), marginal quasi-likelihood (MQL), penalized quasi-likelihood (PQL), pseudo-likelihood (REPL, as implemented in `proc glimmix` in SAS), maximum likelihood via numerical integration (ML) and Bayesian Markov chain Monte Carlo (MCMC).

Analysis of a single dataset by multiple procedures (Preisser *et al.*, 2003) does not necessarily provide much insight into which procedures provide the right answers, and does not cover all aspects of statistical inference. The analytical approach taken for the present study was simulation. Statistical assessments of marginal and random effects procedures for two levels of either binary repeated measures (Masaoud and Stryhn, 2010) or clustered data (Heo and Leon, 2005) are abundant, but these studies do not address the issues related to the additional hierarchical structure.

The objective of the study is to compare marginal and random effects estimation procedures, in terms of statistical properties such as unbiasedness and confidence interval coverage, in a three-level balanced longitudinal design. The comparison includes a range of design parameters at different hierarchical levels. The goal of the comparison is to gain insight into how different estimation approaches deal with the complexity of the design, and to eventually establish some practical guidelines for the choice of statistical procedure for the analysis of balanced, binary repeated measures data with additional hierarchical structure.

2 Statistical Models and Estimation Procedures

Consider binary records y_{ijk} on each of n subjects ($i = 1, \dots, n$) distributed on m clusters ($k = 1, \dots, m$) at t time points ($j = 1, \dots, t$), as well as a set $x_1, \dots, x_p$ of explanatory variables at different hierarchical levels recorded at every time point.

2.1 Random Effects Models

The general form of a logistic random effects repeated measures model (Diggle *et al.*, 2002, Chapter 11) takes the following form:

$$\text{logit}(\Pr(y_{ijk} = 1 | v_k, u_{ijk})) = \beta_0 + \beta_1 x_{1ijk} + \dots + \beta_p x_{pijk} + u_{ijk} + v_k, \quad (2.1)$$

where $(v_1, \dots, v_m)$ are independent random variables with the same distribution and $(u_{i1k}, \dots, u_{itk})$ are a series of autocorrelated random effects with $\rho(u_{ijk}, u_{ij'k}) = \rho^{|j-j'|}$. The most commonly assumed distribution is the Gaussian (normal), say $u_{ijk} \sim N(0, \sigma_2^2)$ where σ_2^2 represents the heterogeneity (variance) between subjects and $v_k \sim N(0, \sigma_3^2)$ where σ_3^2 represents the heterogeneity (variance) between clusters. Model (2.1) is for the conditional probability of an “event” given the random effects v_k and u_{ijk} of the k th cluster and of the i th subject at j th time point, respectively.

A random intercept model arises as a special case of model (2.1) when $\rho = 1$, i.e., the series $(u_{i1k}, \dots, u_{itk})$ of autocorrelated random effects is replaced by the single random effect u_{ik} , assumed $\sim N(0, \sigma_2^2)$, for subject i in cluster k ,

$$\text{logit}(\Pr(y_{ijk} = 1 | v_k, u_{ik})) = \beta_0 + \beta_1 x_{1ijk} + \dots + \beta_p x_{pijk} + u_{ik} + v_k, \quad (2.2)$$

with the same assumptions for (v_k) as above, and the same interpretation of σ_2^2 and σ_3^2 . In our view, model (2.1) forms a better basis for random effects modelling of repeated measures data than the simpler model (2.2) because of its ability to incorporate autocorrelation (Diggle *et al.*, 2002, Chapter 11).

2.2 Random Effects Estimation Procedures

In general, there is no closed form of the full log likelihood function for models (2.1) and (2.2) and numerical procedures are needed to fit the model. Alternatively several approximation algorithms have been proposed, aimed at producing estimates close to the global ML estimate without actually computing the likelihood function (Breslow, 2003). These algorithms carry a number of different names and acronyms typically involving “weighted least squares” and “quasi”- or “pseudo-likelihood”.

Estimation in model (2.2) by numerical approximation most commonly employs the Gauss-Hermite quadrature procedure. Adaptive quadrature (Rabe-Hesketh *et al.*, 2002) is preferable for normally distributed random effects. In adaptive quadrature, the quadrature points are rescaled and shifted to the shape of the log likelihood function. In model (2.1), however, the added random effects at the cluster level pose some challenges for the direct maximization of the log likelihood (ML) and the integration becomes difficult (Diggle *et al.*, 2002) and may substantially increase computation time.

Estimation by Markov chain Monte Carlo (MCMC) techniques in a Bayesian framework, may be viewed as a numerical approach to avoid the computational difficulties of the log likelihood. In this study MCMC techniques are used as an estimation algorithm for the frequentist model rather than for exploring the genuine Bayesian models with informative prior distributions. The MCMC approach has been shown to perform well across a range of settings (Browne and Draper, 2006) and (Masaoud and Stryhn, 2010).

Breslow and Clayton (1993) presented two estimation procedures based on quasi-likelihood function called penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL). Both estimation procedures allow for an extra binomial dispersion parameter (ϕ), see (Skrondal and Rabe-Hesketh, 2007) for discussion of ϕ in PQL. MQL estimates are derived under random effects model assumptions (Goldstein, 1991). Both procedures iteratively employ linear mixed model estimation to an “adjusted” variate obtained by Taylor approximation of the outcome around its current estimated mean, until convergence, using either maximum likelihood (ML) or restricted maximum likelihood (REML), thus results in iterative generalized least squares (IGLS) or restricted iterative generalized least squares (RIGLS), respectively. One major difference between the two algorithms is that MQL does not incorporate the random effects u_i in the linearization of the mean (Molenberghs and Verbeke, 2005, Chapter 9) whereas PQL does. It has been also suggested to refine the approximations by including a second-order term in the Taylor expansions, usually denoted as second order PQL and MQL procedures (Goldstein and Rasbash, 1996) and (Rodriguez and Goldman, 1995). It is well-known that caution should be exercised in using these algorithms because under certain conditions they are prone to bias towards the null (e.g., (Rodriguez and Goldman, 1995) and (Rodriguez and Goldman, 2001)).

In addition, Wolfinger and O’Connell (1993) suggested a similar procedure to PQL, called pseudo-likelihood (PL) procedure. It differs from the quasi-likelihood approach by using a true joint likelihood function in its estimation process. Using either ML or REML in the estimation process results in PL or restricted pseudo-likelihood (REPL), respectively, the REPL procedure allows for both random effects in the linear predictor and correlation structure in the observation scale er-

rors conditional (on the mean) (Wolfinger and O'Connell, 1993). Intuitively, one would expect this procedure to be suitable for models such as model (2.1). Modelling by correlation structure only yields marginal estimates (Molenberghs and Verbeke, 2005, Chapter 22); adding random effects effectively yields a random effects model with serial correlation (Molenberghs and Verbeke, 2005, Chapter 22).

2.3 Marginal Estimation Procedures

The most commonly used procedure to obtain marginal estimates is GEE, generalized estimating equations, which from the onset was devised to deal with repeated measures obtained from multiple subjects (Liang and Zeger, 1986). The terms population-averaged and subject-specific inference originate from this context (Zeger *et al.*, 1988). However, the idea that subjects might be part of a hierarchical structure themselves was not part of the scenario studied. Despite a plethora of extensions of the originally proposed generalized estimating equations (Ziegler *et al.*, 1998), and the availability of manual setup through general purpose software for fitting three-level extended exchangeable correlation structure (Teerenstra *et al.*, 2010) (though it was not included in the present simulation study), to our knowledge no set of estimating equations has been proposed to deal specifically with additional hierarchical structure. Several options can be explored within the classical GEE framework for dealing with one level of clustering of the subjects in addition to the within-subject correlation structure. The simplest idea is perhaps to model clusters by fixed effects (denoted here GEEf) while retaining the usual modelling of within-subject correlation structure. Modelling hierarchical structure by fixed effects has multiple drawbacks, the most important being that it does not allow for inclusion of cluster-level predictors (Dohoo *et al.*, 2009, Chapter 20). An even more crude approach (denoted GEEs), to ignore the additional clustering, would not be expected to yield acceptable cluster-level inference. To achieve correct inference at the cluster level, the GEE handling of correlation structure must be shifted from the subject to the cluster level. This might at first seem to give up on achieving a valid and efficient within-subject inference, but the robustness of GEE procedures to misspecification of working correlation structure should ensure consistency of estimates. The standard choices of GEE working correlation structure do not allow to distinguish between within-cluster and within-subject correlations. In our view the most promising choices of cluster-level correlation structures would be independence (GEEci) and exchangeable (GEEce). Independence correlation structures, effectively corresponding to ordinary logistic regression (OLR) with robust ("sandwich") variance estimates, has been reported to work well for data comprising at least 30 subjects (Ziegler *et al.*, 1998).

An alternative variant of the GEE procedure is alternating logistic regression (ALR). It uses the same estimating equation for the fixed effects as GEE, but differs from GEE by modeling the association among responses (e.g., within subjects) in terms of odds ratios. ALR is numerically more efficient than GEE for large clusters (Carey *et al.*, 1993). The ALR procedure has the advantage of providing standard errors for the association parameters. Furthermore, ALR allows one to distinguish between odds-ratios within clusters and within subclusters (in the current case subjects); however, the within-subject correlation must be modelled as exchangeable. An adapted ALR macro for 3-level of clustering (Kunthel *et al.*, 2014) is available when estimation of the association struc-

ture is of primary interest, though it was not included in our simulation study. For two-level binary repeated measures data, both GEE with an exchangeable correlation structure and ALR yield asymptotically unbiased estimates, which can be nearly efficient relative to GEE with a correctly specified working correlation structure (Masaoud and Stryhn, 2010) and to maximum-likelihood estimates in a fully and correctly specified model (Diggle *et al.*, 2002, Chapter 8).

2.4 Marginal vs. Subject-specific Estimation

The relation between random effects and marginal estimates has been discussed and described (see e.g., (Zeger *et al.*, 1988; Neuhaus, 1992) and (Diggle *et al.*, 2002, Chapter 7)). Zeger *et al.* (1988) provided the conversion formula for logistic regression with normally distributed random effects:

$$\beta^{PA} \approx (c^2 \sigma^2 + 1)^{-1/2} \beta^{SS}, \quad \text{where } c = 16\sqrt{3}/(15\pi) = 0.588. \quad (2.3)$$

For a probit model, replacing the **logit** link function in (2.2) with a **probit** link function Φ^{-1} (Φ being the cumulative distribution function for $N(0, 1)$), the above conversion formula becomes an exact formula (e.g., (McCulloch and Searle, 2008, Chapter 7)):

$$\beta^{PA} = (\sigma^2 + 1)^{-1/2} \beta^{SS}. \quad (2.4)$$

Both formulas can be used to relate subject-specific to population-averaged models/estimates under the assumption that random effects are normally distributed. Without any distributional assumptions on the random effects it holds that the marginal regression parameters are attenuated or diluted (towards zero) relative to the random effects parameters, unless the variance is zero (McCulloch and Searle, 2008, Chapter 7).

3 Simulation Study

The settings for the simulation study were motivated by the scc40 dataset of Dohoo *et al.* (2009, Chapter 31) for repeated measures of somatic cell counts in milk samples from cows housed in multiple herds. In this observational dataset, up to 11 approximately monthly measures were taken on each cow, but missing values of different types occurred. In order to create settings more akin to experimental studies, we consider here balanced and complete series of either $t = 4$ or $t = 8$ measurements per subject. Thirty clusters were included, with 20 subjects per cluster. In the scc40 context, predictors of interest existed at both the herd and cow levels; thus, the simulation design included binary covariates at the cluster and subject levels. Including also (for simplicity) a linear time effect but no interactions with time, the linear predictor included the following parameters set at the indicated true values:

$$\begin{aligned} \beta_0 &= -1 \quad (\text{intercept centered at first time point}), \\ \beta_1 &= .15 \quad (\text{slope for time} = 0, \dots, t-1), \\ \beta_2 &= -1 \quad (\text{coefficient for subject level covariate}), \\ \beta_3 &= 1 \quad (\text{coefficient for cluster level covariate}). \end{aligned}$$

The random part of the model included normally distributed subject and cluster level random effects with standard deviations set at $\sigma_2 = 1.5$ and $\sigma_3 = 0.75$, respectively. These values approximated the estimates in a random intercept model for a binary outcome in the scc40 dataset, as described in section 4. High somatic cell counts are considered an indicator of subclinical mastitis. By the latent variable approximation to the variance partition coefficient (Goldstein *et al.*, 2002), this corresponds to 37% and 9% of the unexplained variance residing at the subject and cluster levels, respectively. Simulated datasets were generated for highly and moderately autocorrelated subject-specific random effects ($\rho = 0.9$ and $\rho = 0.5$) as well as for a random intercept model ($\rho = 1$). Note that the correlation between binary outcomes is different than the correlation between the random effects. In particular, the latent variable approximation with an observation-level variance component of $\pi^2/3$ (Snijders and Bosker, 2012, Chapter 14) yields an intra-class correlation of $\sigma^2/(\sigma^2 + \pi^2/3) = 0.46$, where $\sigma^2 = \sigma_2^2 + \sigma_3^2$, and a first-order correlation of $\rho\sigma^2/(\sigma^2 + \pi^2/3)$, and the values 0.42 and 0.23 for $\rho = 0.9, 0.5$, respectively.

The autocorrelated random effects of each subject were generated by multiplying a vector of t independent variables by the upper triangular factor of the Cholesky decomposition of the correlation matrix (as described in Congdon (2003, Chapter 1)). Generation of the binary outcomes then followed the usual scheme for random effects logistic regression models (Stryhn *et al.*, 2000).

3.1 Software and Settings for Estimation Procedures

The GEE estimation procedure used the implementation of *proc genmod* in SAS software version 9.1, with the following working correlation structures: autoregressive, independence and exchangeable. The ALR estimation procedure used the implementation in SAS with two pairwise odds ratios: within subcluster (subject) odds ratio and within cluster odds ratio. The random effects procedures used the first order MQL and second order PQL procedures, with REML (RIGLS) option and implemented in the *MLwiN* software (version 2.02), the REPL procedure of SAS (*proc glimmix*), as well as the adaptive quadrature algorithms for ML estimation implemented in *Stata* version 10 software (*xtmelogit* command with 7 quadrature points at both the subject and cluster levels). The REPL procedure was set up with cluster and subject-specific random effects and a first order autoregressive correlation structure, and the REML option (Wolfinger and O'Connell, 1993). The REPL procedure included an extra binomial dispersion parameter, and additionally we included versions of the MQL and PQL procedures with an extra binomial dispersion parameter, denoted as MQLx and PQLx. The Bayesian estimation procedures were implemented in *WinBUGS* version 1.4 called from the *R* software using the *R2WinBUGS* package (Sturtz *et al.*, 2005). Vague ("non-informative") prior distributions (i.e. $N(0, 10^6)$) were used for all fixed effects parameters. The uniform distribution for inverse variances, or precisions ($\tau \sim \text{uniform}(0, 100)$) was used (Lambert *et al.*, 2005; Gelman, 2006). The Markov chains were run with 500 burn-in samples (Browne and Draper, 2006), and the subsequent estimates (posterior distribution medians) were based on 2000 samples. These burn-in and estimation sample sizes were arrived at after inspecting MCMC diagnostics for selected datasets.

3.2 Analysis of Results for Simulated Data

The estimates of marginal estimation procedures can be compared either to true subject-specific parameter values or approximate marginal parameter values obtained from the conversion formula (2.3) with $\sigma^2 = \sigma_2^2 + \sigma_3^2$. Inserting the known true variance parameters, yields $\beta^{PA} \approx 0.712 \times \beta^{SS}$. Although the marginal parameters are “theoretical” in the sense that they can only be constructed from the (unknown) variance parameters, they were used to assess the marginal estimation procedures against their expected values. Unless all variances are small, there is little prospect in using marginal estimation procedures to reconstruct the true parameters of random effects models (see (Masaoud and Stryhn, 2010) for discussion of the choice between random effects and marginal estimation procedures). As the fixed effects parameters are on different scales, the results were presented in terms of the relative bias defined as the difference between the average estimate among simulations ($\hat{\beta}$) and the true value (marginal or subject-specific)(β) divided by the true value,

$$\text{relative bias} = \frac{\hat{\beta} - \beta}{\beta} \times 100\% \quad (3.1)$$

The presence of statistically significant bias in the estimates (of both fixed effects and variance parameters) was assessed by a z -test based on the true value and the standard deviation among simulations. The statistical significance of bias in the standard errors was assessed by comparing the mean standard error to a 95% confidence interval for the standard deviation based on the simulations. This simple procedure was considered acceptable because the statistical variation in the estimated standard deviation was generally much larger than that of the mean standard error. Confidence intervals (CIs) were computed by the large-sample normal approximation based on the standard error; for the GEE procedures, the robust standard error was used. The coverage of 95% CIs was computed as the proportion of simulated datasets for which the confidence interval (in the Bayesian analysis: the posterior credibility interval) contained the true parameter.

4 Data Example: High Somatic Cell Counts in Milk Samples

Research based on routine milk samples from dairy herds in programs such as DHI (dairy herd improvement) will almost inevitably involve a repeated measures data structure with additional hierarchical structure (because the cows are located in multiple herds). One possible binary outcome derived from such milk samples is the presence of a high level of somatic cells in the milk sample, e.g., at a somatic cell count threshold of 200 000 cells/ml (De Vliegher *et al.*, 2004). Such a high somatic cell level may be interpreted as an indication of a subclinical (or clinical) mastitis infection (Dohoo and Meek, 1982). We consider here a large dataset collected by Jens Agger and the Danish Cattle Organization in 1993–1994 (Alban and Agger, 1996), of which the scc40 previously mentioned is a subset of 40 herds. We furthermore define high somatic cell counts as the values (strictly) exceeding 200 000 cells/ml. This use of the dataset should not be taken to mean that we consider dichotomizing the outcome as preferable to analyzing the continuous outcome. The dataset contains approximately monthly records throughout one lactation, but for many cows, the series of values throughout the lactation contains missing values. The impact of missing values on the

properties of statistical procedures have been studied in work complementary to the present paper (Masaoud, 2009). For the purpose of illustrating the analysis of complete and balanced repeated measures datasets with additional hierarchical structure, we selected two subdatasets with balanced and complete series of either $t = 4$ or $t = 8$ measurements per cow. Thirty herds were included, with 20 cows per herd. By the observational nature of these data, we defined a binary subject-level predictor as the parity, coded as 1 for heifers and 0 for multiparous cows and a binary cluster-level predictor as the breed, coded as 1 for Holsteins and 0 for other breeds. We additionally included the month of testing (with range either $1 - 4$ or $1 - 8$) as a continuous (linear) term in the model. No interaction terms were included. The two datasets were analyzed by the same statistical procedures as in the main body of the manuscript. The MCMC analysis was based on 100 000 samples from the posterior after a burn-in period of 10 000 samples. No conversion of estimates between SS and PA scales was undertaken because the true model of the data is unknown; therefore, the estimates of marginal and random effects procedures are on different scales.

5 Results

Presentation of results is separated by the type of estimation procedure: based on either marginal or random effects models. Relative biases of estimates and standard errors from the simulation studies are shown in Tables 1–2, coverages of confidence intervals are shown in Figures 1–2. Table 3 gives relative biases of estimates and standard errors for datasets and analyses based on the probit link function. Estimates and standard errors from the analysis of the two somatic cell count subdatasets are presented in Tables 4–5.

5.1 Random Effects Estimation Procedures

As all random effects estimation procedures except REPL are based on a random intercept model, the performance of the procedures in datasets corresponding to this true model ($\rho = 1$) is reviewed first, and subsequently we turn to the results for autoregressive datasets ($\rho < 1$).

5.1.1 Variance Parameters

The two likelihood-based procedures (ML and MCMC) produced fairly accurate variance estimates and standard errors (Table 1). ML estimates of the level 3 variances were slightly attenuated (biased towards zero) and the MCMC standard errors for the same parameter were somewhat inflated (biased away from zero). PQL variance estimates were less accurate, but the direction of the bias depended on whether binomial or extra-binomial dispersion was assumed. In agreement with previous findings in (Masaoud and Stryhn, 2010), PQL showed attenuated estimates, whereas PQLx showed both downward and upward biases (σ_3^2 for $t = 8$, and σ_2^2 for $t = 4$, respectively). The extra-binomial parameter estimates were centered around 0.80 with strongly inflated standard errors. Consequently, PQLx variance estimates were generally higher than PQL estimates. All variance estimates of the REPL procedure were strongly attenuated.

Table 1: Relative bias of estimates and associated standard errors of fixed effects and variance parameters obtained by five random effects estimation procedures, based on analysis of 1000 simulated autocorrelated random effects (with logit link) datasets per setting (t = number of time points, ρ = autocorrelation, parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor), σ_2^2 (variance at subject level), σ_3^2 (variance at cluster level), ϕ (extra binomial dispersion)). Coding for estimation procedures with logit link: PQL (2nd order penalized quasi-likelihood), PQLx (2nd order penalized quasi-likelihood with extra binomial dispersion), REPL (restricted pseudo-likelihood), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

t	ρ	Par.	Random effects estimation procedures									
			PQL		PQLx		REPL		ML		MCMC	
			Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
8	1	β_0	0.0	-5.3*	2.7 [‡]	-5.9*	-8.5 [‡]	-2.6	0.9	-4.9	0.6	-5.5*
		β_1	0.9 [†]	-2.4*	2.7 [‡]	-11.8*	-6.4 [‡]	-8.6*	0.7 [†]	-0.2	0.9 [†]	-0.6
		β_2	-3.0 [‡]	1.1	-0.6	1.3	-11.4 [‡]	0.2	0.0	1.0	0.2	0.7
		β_3	-1.9	-5.8*	0.6	-5.8*	-11.0 [‡]	-2.3	1.0	-5.7*	1.0	-4.8*
		σ_2^2	-8.6 [‡]	-8.9*	2.7 [‡]	-14.7*	-18.8 [‡]	-6.4*	-0.1	0.3	1.0 [‡]	0.0
		σ_3^2	-11.5 [‡]	-2.1	-7.0 [‡]	-2.2	-19.6 [‡]	2.1	-6.2 [‡]	-0.8	3.4 [‡]	12.7*
		ϕ			-16.2 [‡]	44.9*	-21.4 [‡]	20.9*				
	9	β_0	-9.4 [‡]	-5.8*	-7.6 [‡]	-6.3*	-18.6 [‡]	-2.6	-9.5 [‡]	-5.2*	-9.2 [‡]	-4.7*
		β_1	-8.6 [‡]	-14.2*	-6.7 [‡]	-20.2*	-17.3 [‡]	-11.4*	-9.1 [‡]	-11.7*	-8.9 [‡]	-12.1*
		β_2	-10.4 [‡]	-2.2	-8.7 [‡]	-1.8	-19.7 [‡]	-1.1	-9.4 [‡]	-0.5	-9.3 [‡]	-0.3
		β_3	-9.7 [‡]	-6.0*	-7.9 [‡]	-6.0*	-19.3 [‡]	-2.8	-8.8 [‡]	-5.6*	-8.4 [‡]	-4.0*
		σ_2^2	-43.1 [‡]	-10.9*	-36.2 [‡]	-15.4*	-54.6 [‡]	-2.8	-39.6 [‡]	2.5	-39.0 [‡]	2.9
		σ_3^2	-26.1 [‡]	-1.2	-23.1 [‡]	-1.1	-35.7 [‡]	4.6	-24.7 [‡]	1.4	-16.9 [‡]	16.2*
		ϕ			-12.1 [‡]	108.8*	-15.8 [‡]	60.3				
	.5	β_0	-22.6 [‡]	-8.6*	-21.9 [‡]	-9.0*	-28.3 [‡]	-4.6*	-22.9 [‡]	-7.8*	-23.1 [‡]	-6.2*
		β_1	-22.6 [‡]	-15.0*	-22.0 [‡]	-17.8*	-27.3 [‡]	-9.0*	-22.9 [‡]	-12.9*	-23.0 [‡]	-13.5*
		β_2	-23.6 [‡]	-3.0	-23.0 [‡]	-3.1	-28.4 [‡]	-0.5	-23.9 [‡]	-0.4	-23.8 [‡]	-0.6
		β_3	-22.9 [‡]	-7.8*	-22.2 [‡]	-7.9*	-27.8 [‡]	-4.0	-23.1 [‡]	-7.4*	-23.1 [‡]	-5.4*
		σ_2^2	-88.3 [‡]	-11.6*	-86.4 [‡]	-17.1*	-93.9 [‡]	-7.6*	-87.4 [‡]	-1.7	-87.6 [‡]	-5.6*
		σ_3^2	-45.5 [‡]	0.3	-44.6 [‡]	0.1	-47.6 [‡]	5.0*	-45.9 [‡]	3.2	-40.4 [‡]	19.3*
		ϕ			-5.4 [‡]	187.3*	-4.4 [‡]	61.9*				
4	1	β_0	-4.5 [‡]	-5.1*	2.1 [†]	-6.0*	-14.8 [‡]	-4.2*	-1.0	-3.6	-0.8	-2.8
		β_1	-2.5 [‡]	0.5	5.8 [‡]	-14.9*	-10.9 [‡]	-8.2*	-0.5	2.7	-0.2	2.3
		β_2	-4.7 [‡]	-0.3	1.1 [†]	1.6	-15.5 [‡]	-1.1	-0.2	1.1	0.2	1.2
		β_3	-5.4 [‡]	-5.6*	0.4	-5.2*	-16.3 [‡]	-3.1	-1.2	-4.6*	-0.9	-2.5
		σ_2^2	-14.2 [‡]	-13.8*	17.9 [‡]	-23.3*	-22.2 [‡]	-12.2*	1.2 [‡]	2.7	3.1 [‡]	2.7
		σ_3^2	-13.8 [‡]	-1.6	-2.5 [†]	-1.1	-25.9 [‡]	1.9	-5.9 [‡]	0.9	3.9 [‡]	14.1*
		ϕ			-23.0 [‡]	19.4*	-27.7 [‡]	11.8*				
	9	β_0	-8.8 [‡]	-5.1*	-3.2 [‡]	-6.3*	-20.0 [‡]	-3.2	-6.8 [‡]	-3.5	-6.6 [‡]	-1.4
		β_1	-6.7 [‡]	-6.9*	-0.1	-18.9*	-16.4 [‡]	-10.1*	-5.6 [‡]	-4.2	-5.1 [‡]	-4.8*
		β_2	-8.7 [‡]	-1.7	-3.5 [‡]	-0.4	-20.0 [‡]	-0.6	-6.0 [‡]	1.3	-5.7 [‡]	1.3
		β_3	-9.3 [‡]	-6.2*	-4.1 [‡]	-6.2*	-21.0 [‡]	-2.9	-6.8 [‡]	-5.3*	-6.4 [‡]	-2.9
		σ_2^2	-35.5 [‡]	-18.0*	-10.7 [‡]	-28.4*	-45.3 [‡]	-11.7*	-25.4 [‡]	1.7	-24.1 [‡]	0.5
		σ_3^2	-21.9 [‡]	-4.7*	-12.6 [‡]	-4.8*	-34.6 [‡]	0.0	-17.6 [‡]	-1.1	-9.2 [‡]	13.1*
		ϕ			-19.5 [‡]	37.3*	-22.7 [‡]	20.9*				
	.5	β_0	-21.7 [‡]	-5.9*	-18.8 [‡]	-7.2*	-27.8 [‡]	-1.8	-21.4 [‡]	-4.3*	-21.5 [‡]	-1.4
		β_1	-20.5 [‡]	-7.5*	-17.5 [‡]	-14.4*	-26.4 [‡]	-4.6*	-20.3 [‡]	-5.0*	-20.1 [‡]	-5.2
		β_2	-20.3 [‡]	-1.8	-17.4 [‡]	-2.6	-26.9 [‡]	1.6	-19.8 [‡]	3.7	-19.7 [‡]	3.0
		β_3	-21.6 [‡]	-5.6*	-18.7 [‡]	-6.0*	-28.1 [‡]	-0.9	-21.2 [‡]	-4.6*	-21.3 [‡]	-0.6
		σ_2^2	-80.8 [‡]	-15.7*	-71.6 [‡]	-31.0*	-85.7 [‡]	-2.7	-77.1 [‡]	0.7	-77.4 [‡]	-6.1*
		σ_3^2	-41.3 [‡]	-5.5*	-36.9 [‡]	-6.7*	-45.8 [‡]	3.6	-40.6 [‡]	-0.6	-34.5 [‡]	13.8*
		ϕ			-10.4 [‡]	87.4*	-9.7 [‡]	25.9*				

[†] significant bias in estimate at $P < 0.05$; [‡] significant bias in estimate at $P < 0.01$;

* significant bias in standard error at $P < 0.05$

Due to the scaling by the variance parameters inherent in random effects estimation procedures, any bias in estimated variances is likely to affect the fixed effects as well (and in the same direction), in particular for moderate to large variance components. For example, all fixed effects estimates for REPL indeed showed substantial bias (range 6–16%) towards zero. Also, the attenuation of both fixed effects and variance parameters for PQL estimation was more pronounced for shorter time series, corresponding to less replication at the subject level (Masaoud and Stryhn, 2010).

5.1.2 Level 3 Parameters

A similar qualitative behaviour was expected for the intercept and the predictor at the highest (cluster) level, and generally the results confirmed this. Likelihood-based procedures gave unbiased estimates but in some cases slightly underestimated the standard error. PQL procedures showed the same bias in the standard error, and some instances of minor negative (PQL) or positive (PQLx) bias in the estimates. As noted above, REPL estimates were clearly biased towards zero. CI coverage (Figure 1) was close to or slightly below nominal for all procedures except REPL.

5.1.3 Level 1 and 2 Parameters

The subject-level parameter was estimated without any biases for ML, MCMC and PQLx; a small negative bias (3–5%) was present in PQL estimates. The regression coefficient for time (level 1) was estimated without bias by ML, MCMC and PQL (except for $t = 4$), whereas the PQLx estimates were moderately inflated, had too small standard errors and showed undercoverage of CIs.

5.1.4 Autocorrelated Data ($\rho < 1$)

For all random effects procedures, estimates of both fixed effects and variance parameters were attenuated in autoregressive datasets (Table 1). In the vast majority of settings, the biases were statistically significant. The relative bias was strongest for the variance parameters, increased markedly from $\rho = 0.9$ to $\rho = 0.5$, and was also somewhat larger for the long time series ($t = 8$). Standard errors were also clearly underestimated (up to 18%) in several cases, including the variance parameters and in the long series also the time effect. The bias in standard errors was less severe for the short series ($t = 4$). Contrary to the other procedures, MCMC estimation produced inflated (up to 19%) estimates for σ_3^2 . Generally, the setting least affected was $(\rho, t) = (0.9, 4)$ where biases for fixed parameter estimates of ML, MCMC and PQL were below 10%, and CI coverage was above 90%. In other settings, the biases for these procedures ranged up to 24% for fixed effects and 88% for variance parameters, and CI coverage could go below 50% (for $(\rho, t) = (0.5, 8)$).

Adding the extra-binomial dispersion parameter to the PQL procedure did not alleviate the attenuation of PQL estimates substantially. For low autocorrelation ($\rho = 0.5$), both PQL procedures and the likelihood-based procedures showed downward biases and undercoverages of similar magnitudes. However, the extra-dispersion parameter tended to increase in value (show less bias) for decreasing ρ . Although the only random effects procedure examined which incorporated an autoregressive parameter, the REPL procedure was equally affected by the autocorrelation as the other

procedures. Moreover, for both values of ρ the REPL estimates had stronger bias and lower CI coverage than those of the other procedures.

5.2 Marginal Estimation Procedures

Generally, estimates from OLR, ALR, MQL, and the GEE procedures except GEEf agreed closely (Table 2), and all showed a small negative relative bias in the range 3–6%. For estimation of a marginal parameter, all the methods are asymptotically consistent, and it is plausible that this bias is due to approximation error in the calculation of the marginal true value; see Section 5.2.3 below. As the asymptotic consistency does not rely on assumptions about the true correlation structure, the presence of autocorrelation ($\rho < 1$) in the data was expected to have less of an impact for marginal than random effects procedures.

5.2.1 Level 3 Parameters

The fixed cluster effects included in the linear predictor for GEEf precluded estimation of effects at the cluster level. For the other procedures, the relative bias remained relatively constant in the 3–6% range irrespective of the autocorrelation (ρ); thus, the differences in performance were essentially in the standard errors. GEEce estimates for the cluster level covariate appear to be numerically off the estimates from the other procedures. As expected, OLR grossly underestimated the standard errors at the cluster level. Strongly underestimated standard errors were also seen for the GEEs procedure, a consequence of the lack of cluster level effects in the (variance) estimating equations. The other procedures generally showed a minor (up to 7.5%) downward relative bias in the standard errors, and CI coverage at or moderately below the nominal level (Figure 2), irrespective of the value of ρ .

5.2.2 Level 1 and 2 Parameters

Estimates from GEEf showed a small upwards bias (less than 5%), in the opposite direction of the downwards bias displayed by all other procedures. Standard errors were generally, except for some settings for the time coefficient, unbiased for ALR and all GEE procedures excluding GEEs, which had substantial upwards bias for the level 2 coefficient. Both the MQL and MQLx procedures showed biases in either direction for the time coefficient. Stronger biases were noted for OLR, also in both directions. CI coverages varied around the nominal level for all procedures except OLR.

5.2.3 Overall Marginal Bias

We examined the small negative bias experienced by almost all marginal estimation procedures by rerunning the simulation study using a probit model both to generate the data and fit the models involved in the estimation procedures. As previously noted for a probit model, the conversion formula (2.4) is exact. Table 3 gives the relative bias for a subset of the previously considered estimation procedures. The OLR, GEEce, GEEci and ALR procedures gave virtually unbiased estimates, although some bias in the standard errors remained (similar to the previously discussed results in Table 2). As the exact conversion formula thus produced unbiased estimates, we consider the approximation

Table 2: Relative bias of fixed effects parameter estimates (against marginal true values) and their standard errors obtained by eight marginal estimation procedures, based on analysis of 1000 simulated autocorrelated random effects (with logit link) datasets per setting (t = number of time points, ρ = autocorrelation, parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor)). Coding for estimation procedures with logit link: OLR (ordinary logistic regression), GEEci (generalized estimating equations (GEE) with independence correlation at cluster level), GEEe (GEE with exchangeable correlation at cluster level), GEEf (GEE with fixed effects for cluster level and autoregressive correlation at subject level), GEEs (GEE with autoregressive correlation at subject level), ALR (alternating logistic regression), MQL (marginal quasi-likelihood), MQLx (marginal quasi-likelihood with extra binomial dispersion).

t	ρ	Par.	Marginal estimation procedures											
			OLR			GEEci			GEEe			GEEf		
			Est.	SE	SE	Est.	SE	SE	Est.	SE	SE	Est.	SE	SE
8	1	β_0	-3.5 [†]	-58.5*	-6.4*	-4.0 [†]	-6.5*	n/a	n/a	n/a	-3.7 [†]	-38.8*	-3.6 [†]	-6.3*
		β_1	-3.9 [†]	18.8*	-2.7	-3.9 [†]	-2.8	4.2 [†]	-3.3	-3.5 [†]	-3.3	-3.9 [†]	-2.8	-3.9 [†]
		β_2	-4.7 [†]	-38.7*	0.0	-4.7 [†]	0.0	3.1 [†]	1.6	-4.5 [†]	12.0*	-4.8 [†]	0.0	-4.7 [†]
		β_3	-3.6 [†]	-72.3*	-5.5*	-3.9 [†]	-5.3*	n/a	n/a	-3.0 [†]	-50.6*	-3.6 [†]	-5.7*	-3.5 [†]
9		β_0	-4.2 [†]	-57.0*	-5.5*	-4.5 [†]	-4.9*	n/a	n/a	-4.2 [†]	-40.3*	-4.2 [†]	-5.4*	-4.2 [†]
		β_1	-3.8 [†]	1.2	-2.8	-3.8 [†]	-2.8	3.4 [†]	1.4	-3.9 [†]	-2.0	-3.8 [†]	-2.8	-3.8 [†]
		β_2	-4.5 [†]	-34.1*	-2.4	-4.5 [†]	-2.4	2.8 [†]	3.2	-4.5 [†]	8.5*	-4.5 [†]	-2.5	-4.5 [†]
		β_3	-3.8 [†]	-71.4*	-7.0*	-4.0 [†]	-5.2*	n/a	n/a	-3.7 [†]	-53.0*	-3.9 [†]	-5.5*	-3.8 [†]
5		β_0	-3.7 [†]	-56.1*	-7.5*	-5.0 [†]	-7.7*	n/a	n/a	-4.6 [†]	-46.9*	-3.7 [†]	-7.5*	-4.3 [†]
		β_1	-3.7 [†]	-8.9*	-6.3*	-4.0 [†]	-8.3*	3.0 [†]	-3.0	-3.9 [†]	-6.3*	-3.7 [†]	-6.3*	-3.7 [†]
		β_2	-4.9 [†]	-14.0*	-3.5	-5.3 [†]	-3.1	1.7 [†]	-2.8	-5.3 [†]	14.4*	-4.9 [†]	-3.6	-4.9 [†]
		β_3	-4.1 [†]	-70.6*	-5.6*	-4.4 [†]	-6.8*	n/a	n/a	-4.1 [†]	-61.3*	-4.1 [†]	-7.0*	-4.1 [†]
4	1	β_0	-5.6 [†]	-46.4*	-5.1*	-5.9 [†]	-5.3*	n/a	n/a	-5.7 [†]	-35.0*	-5.6 [†]	-5.1*	-5.6 [†]
		β_1	-5.4 [†]	23.9*	-0.1	-5.0 [†]	-0.1	3.1 [†]	2.1	-5.0 [†]	2.3	-5.0 [†]	-0.2	-5.0 [†]
		β_2	-4.8 [†]	-22.9*	-1.0	-4.8 [†]	-1.0	3.3 [†]	-1.7	-4.7 [†]	7.3*	-4.8 [†]	-1.0	-4.7 [†]
		β_3	-5.8 [†]	-61.6*	-5.5*	-6.0 [†]	-5.5*	n/a	n/a	-5.8 [†]	-46.2*	-5.8 [†]	-5.7*	-5.7 [†]
9		β_0	-5.2 [†]	-45.3*	-4.2*	-5.6 [†]	-4.0	n/a	n/a	-5.4 [†]	-34.9*	-5.2 [†]	-4.1*	-5.2 [†]
		β_1	-3.7 [†]	11.5*	-4.3*	-3.7 [†]	-4.3*	4.1 [†]	-2.5	-3.9 [†]	-2.7	-3.7 [†]	-4.3*	-3.7 [†]
		β_2	-4.3 [†]	-20.0*	-1.0	-4.3 [†]	-1.0	3.7 [†]	-1.8	-4.2 [†]	7.5*	-4.3 [†]	-0.9	-4.2 [†]
		β_3	-5.3 [†]	-61.0*	-4.9*	-5.6 [†]	-5.4*	n/a	n/a	-5.3 [†]	-47.3*	-5.3 [†]	-5.5*	-5.2 [†]
5		β_0	-5.8 [†]	-44.0*	-4.6*	-6.2 [†]	-4.4*	n/a	n/a	-5.8 [†]	-37.4*	-5.8 [†]	-4.6*	-5.8 [†]
		β_1	-4.3 [†]	2.9	-3.4	-4.3 [†]	-3.4	3.1 [†]	-0.4	-4.3 [†]	-0.4	-4.3 [†]	-3.4	-4.3 [†]
		β_2	-3.8 [†]	-6.6*	1.1	-3.9 [†]	1.1	3.6 [†]	0.3	-3.8 [†]	11.4*	-3.8 [†]	1.1	-3.8 [†]
		β_3	-5.6 [†]	-59.4*	-4.7*	-5.9 [†]	-4.7*	n/a	n/a	-5.6 [†]	-51.5*	-5.6 [†]	-4.9*	-5.5 [†]

[†] significant bias in estimate at $P < 0.05$; * significant bias in estimate at $P < 0.01$; † significant bias in standard error at $P < 0.05$

Table 3: Relative bias of fixed effects parameter estimates (against marginal true values) and their standard errors obtained by eight marginal estimation procedures, based on analysis of 1000 simulated autocorrelated random effects (with **probit** link) datasets per setting (t = number of time points, ρ = autocorrelation, parameters: β_0 (intercept), β_1 (time coefficient), β_2 (subject level factor), β_3 (cluster level factor))). Coding for estimation procedures with **probit** link: OLR (ordinary logistic regression), GEEci (generalized estimating equations (GEE) with independence correlation at cluster level), GEEce (GEE with exchangeable correlation at cluster level), GEEf (GEE with fixed effects for cluster level and autoregressive correlation at subject level), GEEs (GEE with autoregressive correlation at subject level), ALR (alternating logistic regression).

t	ρ	par.	Marginal estimation procedure (probit link)											
			OLR		GEEci		GEEce		GEEf		GEEs		ALR	
			Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
8	1	β_0	1.7 [†]	-64.8*	1.7 [†]	-2.0	1.7	-2.7	n/a	n/a	1.7 [†]	-38.8*	1.8 [†]	-1.9
		β_1	1.4 [†]	30.6*	1.4 [†]	-3.2	1.2 [†]	-3.2	11.6 [†]	-1.7	1.3 [†]	-3.0	1.2 [†]	-3.4
		β_2	0.8	-46.8*	0.8	2.2	0.7	2.4	11.2 [†]	1.3	0.8	12.2*	0.8	2.4
		β_3	2.4 [†]	-77.6*	2.4 [†]	-8.1*	2.4 [†]	-8.4*	n/a	n/a	2.3 [†]	-52.0*	2.4 [†]	-8.2*
9	0	β_0	-0.6	-64.3*	-0.6	-6.8*	-0.3	-7.5*	n/a	n/a	-0.6	-44.2*	-0.5	-7.0*
		β_1	0.6	1.3	0.6	-0.8	0.6	-0.6	10.3 [†]	1.8	0.6	-0.5	0.6	-0.8
		β_2	1.4 [†]	-44.0*	1.4 [†]	-3.5	1.4 [†]	-3.5	11.2 [†]	-5.7*	1.4 [†]	6.6*	1.4 [†]	-3.6
		β_3	-0.5	-75.3*	-0.5	-1.8	-0.2	-2.3	n/a	n/a	-0.5	-53.1*	-0.5	-2.0
5	0.3	β_0	0.3	-61.1*	0.3	-3.4	0.2	-3.0	n/a	n/a	0.3	-48.9*	0.4	-3.6
		β_1	-0.5	-9.7*	-0.5	-4.0	-0.5	-3.9	8.2 [†]	-0.5	-0.5	-1.4	-0.5	-4.0
		β_2	0.3	-20.3*	0.3	-3.1	0.3	-3.0	9.0 [†]	-3.0	0.3	16.5*	0.3	-3.2
		β_3	1.9 [†]	-75.0*	1.9 [†]	-5.0*	1.8	-5.2*	n/a	n/a	1.9 [†]	-63.5*	1.9 [†]	-5.1*
4	1	β_0	-0.7	-54.0*	-0.7	-5.9*	-0.7	-6.3*	n/a	n/a	-0.7	-38.8*	-0.7	-5.9*
		β_1	-0.1	45.4*	-0.1	-0.6	-0.1 [†]	-0.6	10.9 [†]	2.2	-0.1 [†]	2.1	-0.1	-0.9
		β_2	0.5	-32.3*	0.5	-1.5	0.5	-1.4	11.4 [†]	-1.8	0.7	7.7*	0.5	-1.4
		β_3	-0.6	-67.9*	-0.6	-5.8*	-0.6	-5.7*	n/a	n/a	-0.6	-48.6*	-0.6	-5.9*
9	0	β_0	-0.4	-51.8*	-0.4	-2.7	-0.4	-2.7	n/a	n/a	-0.6	-37.7*	-0.4	-2.8
		β_1	0.9	24.7*	0.9	-1.9	0.9	-1.9	11.2 [†]	0.5	0.7	0.2	0.9	-1.9
		β_2	0.9 [†]	-25.2*	0.9 [†]	2.5	0.9 [†]	2.6	11.3 [†]	3.1	1.0 [†]	13.3*	0.9 [†]	2.6
		β_3	-1.0	-67.6*	-1.0	-6.8*	-0.9	-6.4*	n/a	n/a	-1.0	-50.8*	-1.0	-6.9*
5	0.7	β_0	0.7	-51.5*	0.7	-5.7*	0.8	-5.6*	n/a	n/a	0.6	-43.1*	0.7	-5.7*
		β_1	1.1	6.7*	1.1	-0.4	1.1	-0.4	10.6 [†]	2.4	0.9	2.5	1.1	-0.3
		β_2	0.5	-13.6*	0.5	-1.9	0.5	-1.9	10.0 [†]	-2.1	0.6	10.8*	0.5	-1.9
		β_3	0.8	-66.3*	0.8	-6.6*	0.9	-6.6*	n/a	n/a	0.9	-56.8*	0.9	-6.7*

[†] significant bias in estimate at $P < 0.05$; [†] significant bias in estimate at $P < 0.01$; * significant bias in se at $P < 0.05$

formula as a plausible source of the general minor downward bias seen in Table 2. The results for the probit link also confirmed the suspected positive bias in GEEf estimates. As before, the autocorrelation had only little impact on the performance of the estimation procedures.

5.3 High Somatic Cell Count Data

The estimates from analysis of the two datasets (with $t = 4$ and $t = 8$) were remarkably different. Subject matter conclusions should clearly not be based on analyses of purposively selected subdatasets because of any inherent biases in the selection procedure. For the comparison of the statistical procedures, the difference in the estimates is less of a problem, and does on the other hand permit a comparison across a range of parameter values.

Within each of the datasets, the estimates from random effects models varied substantially (Table 4). The numerically smallest estimates were obtained from REPL for all parameters, whereas the numerically largest estimates were mostly from ML and MCMC (which in turn agreed fairly closely). The ratios between smallest and largest estimates were in the range of 70–86% for fixed effects parameters, and around 57% for variance parameters. The variance estimates were largest for ML and MCMC, less attenuated for PQLx than PQL and generally lowest for REPL. The differences between fixed effects estimates were to a large degree attributable to the different variance estimates and the different scales for fixed effects resulting hereof. Indeed, after conversion to a common marginal scale using the conversion formula (2.3) and the estimated variances, the differences between fixed effects estimates were considerably smaller (Appendix A, Table A1). This scale-dependence occurs whenever the random effects variances are large, as for the present data. On marginal scale, the REPL estimates and standard errors were closer to zero, except for the time coefficient (β_1), but otherwise the estimates were mostly close without any particular patterns.

Generally, estimates from OLR, ALR, MQL, and the GEE procedures agreed fairly closely (Table 5), except for the GEEce estimates at the cluster level (β_0, β_3) being numerically larger than for other procedures. There was also quite good agreement with the converted estimates from random effects procedures (except REPL). The converted estimates seemed less variable than the different marginal estimates, possibly a consequence of all these models having the same assumptions about correlation structure, whereas the correlation structures differed among the GEE procedures. The standard errors for the time coefficient were similar across all marginal procedures (except OLR), but for the other parameters there was better agreement between GEEci, GEEce and ALR than the rest.

Table 4: Parameter estimates and associated standard errors obtained by different statistical procedures for two somatic cell count datasets comprising 20 cows in each of 30 herds and measures at $t = 4$ or $t = 8$ time points, respectively. Parameters (see text for model details): β_0 (intercept), β_1 (month), β_2 (parity), β_3 (breed), σ_2^2 (between-cow variance), σ_3^2 (between-herd variance), ϕ (extra binomial dispersion). Coding for estimation procedures with logit link: PQL (2nd order penalized quasi-likelihood), PQLx (2nd order penalized quasi-likelihood with extra binomial dispersion), REPL (restricted pseudo-likelihood), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

		Estimation procedures									
		PQL		PQLx		REPL		ML		MCMC	
t	Par.	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
4	β_0	-2.232	.348	-2.456	.353	-2.126	.267	-2.621	.377	-2.673	.410
	β_1	.383	.061	.451	.048	.368	.041	.404	.063	.411	.064
	β_2	-1.212	.286	-1.287	.304	-1.006	.213	-1.367	.296	-1.369	.295
	β_3	.119	.393	.146	.417	.115	.312	.165	.413	.159	.438
	σ_2^2	5.925	.648	7.424	.726	4.670	.446	7.864	1.093	8.462	1.275
	σ_3^2	.557	.301	.625	.342	.395	.196	.655	.332	.686	.420
	ϕ			.547	.018	.474	.021				
8	β_0	-1.813	.306	-1.878	.311	-1.624	.252	-1.951	.320	-1.967	.345
	β_1	.217	.020	.227	.018	.195	.017	.222	.020	.221	.021
	β_2	-1.229	.218	-1.263	.233	-1.035	.173	-1.362	.228	-1.364	.224
	β_3	-.160	.385	-.159	.394	-.118	.316	-.159	.399	-.132	.432
	σ_2^2	4.480	.395	4.961	.418	3.224	.287	5.542	.587	5.693	.608
	σ_3^2	.773	.292	.812	.309	.546	.208	.839	.321	.931	.402
	ϕ			.744	.016	.641	.017				

6 Discussion

6.1 Random Effects Estimation Procedures

In data generated by random intercept models ($\rho = 1$), the ML and MCMC procedures performed well and had fairly similar estimation errors. Generally, the estimation bias in PQL is known (McCulloch and Searle, 2008, Chapter 14) and could be partly due to the well-known attenuation of variance parameters by PQL in certain settings (Evans *et al.*, 2001). However, the results (especially when $\phi = 1$) were in accordance with the those of Molenberghs and Verbeke (2005, Chapter 14), who reported that the performance of PQL could be improved by increasing the number of subjects. Venables and Ripley (2002) concluded that allowing for ϕ in certain applications, yielded regression estimates that were closer to the maximum likelihood estimates. However, Skrondal and Rabe-Hesketh (2007) argue that ϕ is redundant parameter when using quasi-likelihood methods for modelling binary data. Our results confirm that ϕ is redundant parameter in such case. Furthermore it showed an evidence of inflated standard errors for ϕ , which does not support the suggestion by

[illegible]

Figure 1: Confidence interval coverage for estimates of fixed effects parameters of different estimation procedures, based on 1000 simulated autocorrelated random effects (with logit link) datasets per setting ($t = \text{number of time points}$, $(\rho = (1, .9, .5) \sim (\square, \triangle, \circ))$). Coding for estimation procedures: PQL (2nd order penalized quasi-likelihood), PQLx (2nd order penalized quasi-likelihood with extra binomial dispersion), REPL (restricted pseudo-likelihood), MCMC (Bayesian Markov chain Monte Carlo), ML (maximum likelihood).

Figure 2: Confidence interval coverage for estimates of fixed effects parameters of seven marginal estimation procedures, based on 1000 simulated autocorrelated random effects (with logit link) datasets per setting (t = number of time points, ($\rho = (1, .9, .5) \sim (\square, \triangle, \circ)$)). Coding for estimation procedures: OLR (ordinary logistic regression), GEEf (generalized estimating equations (GEE) with fixed effects for cluster level and autoregressive correlation at subject level), GEEce (GEE with exchangeable correlation at cluster level), GEEci (GEE with independence correlation at cluster level), GEEs (GEE with autoregressive correlation at subject level), ALR (alternating logistic regression), MQL (marginal quasi-likelihood).

Yang *et al.* (2000) of using ϕ as a diagnostic tool, and by Barbosa and Goldstein (2000) to “allow for model misspecification”. Heo and Leon (2005) concluded that the full likelihood approach “appears to be preferable for the analysis of clustered binary observations with underlying random effects models”.

In datasets with autocorrelation ($\rho < 1$), the downward biased estimates can be seen largely as a scaling effect caused by the underestimation of the random effect variances. Inference for all fixed effects parameters was strongly affected, and the biases of estimates were of similar magnitude. The associated pattern of increases in the inflated standard errors for ϕ with decreases in the correlation value, could be due to the increase of the variability within each subject as the correlation decreases. This finding may raise some concern about the approximation procedures to the log likelihood. However, more research is needed to confirm it, especially in simpler settings such as correlated binary data.

The REPL estimation procedure performed poorly and produced unsatisfactory estimates regardless of autocorrelation ($\rho = 0.5, 0.9, 1$). These findings are in support of those by Evans *et al.* (2001) for the variance component and by (Masaoud and Stryhn, 2010) for regression estimates. This could be due to the inclusion of both random effects and a correlation structure in the REPL procedure, and thus modelling parts of the variance/correlation structure on different scales (Molenberghs and Verbeke, 2005, Chapter 22). The ability of the REPL procedure to incorporate autocorrelation did not render the estimates less susceptible to attenuation bias in datasets with autocorrelation than the other random effects procedures.

6.2 Marginal Estimation Procedures

The results suggest that the small general bias observed for marginal estimation procedures could be due to the conversion formula (2.3). However, to our knowledge, the assessment of the accuracy of this formula in practice has not been reported yet. This conversion formula relies on the variance of normally distributed random effects to scale subject-specific to population-averaged models/estimates. The avoidance of such scaling problems by separating fixed and random effects estimates was one of the key ideas behind the development of marginalized models (Heagerty and Zeger, 2000). The application of marginalized models to repeated measures data with additional hierarchical structure has to our knowledge not been reported.

Marginal estimates were little affected by the presence of autocorrelation; similar performances of the different procedures were seen in all settings ($\rho = 0.5, 0.9, 1$). ALR, GEEci, and GEEce performed fairly well with only a few instances of minor statistically significant bias. MQL (apart from the fluctuation in the standard error for the time coefficient) performed on a par with ALR, and in agreement with previous finding in (Masaoud and Stryhn, 2010). The results indicate that accounting for clustering at the highest hierarchical level is sufficient (GEEce, GEEci). In fact, standard errors for these procedures and ALR showed a larger bias at the cluster level than at levels below, probably an effect of the relatively low number of clusters (Ziegler *et al.*, 1998). This is an interesting result that we think needs more research to validate it in different settings and designs of binary data. Furthermore, we suggest that further research may include the relatively recently adapted ALR macro for 3-level of clustering (Kunthel *et al.*, 2014).

Accounting for the additional hierarchical structure by fixed effects (GEEf) resulted in biased estimates. However, ignoring the hierarchical structure in the data in (OLR and GEEs) resulted in inflated and biased standard error of the cluster level fixed effect, this is in agreement with those reported by Diggle *et al.* (2002, Chapter 7) and Dohoo and Stryhn (2006). Allowing for ϕ in MQL had almost no impact on the regression estimates of fixed effects. An explanation could be that MQL does not incorporate the random effects in the linearization of the mean (Molenberghs and Verbeke, 2005, Chapter 14).

6.3 High Somatic Cell Count Data

The analysis of the two sample datasets illustrated the difficulties with comparing results of different estimation procedures from analysis of a single dataset. The scale-dependence of fixed effects estimates of random effects models may render any direct comparisons between these estimates pretty meaningless, in particular if the estimated variance components are large. Converting the estimates to marginal scale does not help with determining the “best” estimates in the random effects model, if these were of primary interest. In that case, we would generally recommend to use the results from the estimation procedure with least biases for the variance components, as discussed above.

The analysis of the two sample datasets also illustrated some features of the findings from the simulation studies (e.g., the intermediate character of REPL estimates, the biases of certain variance estimates, the sufficiency of accounting for clustering at the highest hierarchical level in (GEEce, GEEci)). On the other hand, some results were unexpected from the simulation study, such as the deviation of the GEEce method in the cluster-level parameters. This effect would seem to be due to sampling variability, and thus once more illustrates the difficulty with drawing conclusions about the performance of procedures from analyses of single dataset.

The dataset also illustrates that in multi-level hierarchical structures, the choice between population-averaged and subject-specific interpretations of regression coefficients can be less clearcut than with a single level of clustering. Because modern dairy herds are essentially limited to a single breed, a population-averaged interpretation for breed effects across both cows and herds is most natural. The parity of a cow can however be interpreted both as subject-specific for a heifer within a specific herd (with the parity changing in the subsequent lactation), and as population-averaged across cows (in particular when no they are longer heifers) but within a specific herd, and population-averaged across both cows and herds. Such nuances require a good understanding of these concepts and of the appropriate analytical tools to achieve estimates with a desired interpretation.

6.4 Recommendations

We conclude with a discussion of the implications of the current findings for the choice of procedure. For the choice between marginal and random effects approaches, this study adds only little to existing knowledge (Masaoud and Stryhn, 2010). The bias seen in marginal estimates should be of no concern for the use of marginal procedures if the interpretation in this study is correct that it is caused by the conversion formula. For autoregressive data, the random effects procedures performed

poorly (as was found also in (Masaoud and Stryhn, 2010)), therefore marginal procedures may seem more attractive, unless the time series is very short (less than 4 time points).

Generally, the likelihood-based random effects procedures (ML, MCMC) performed better than methods based on quasi-or pseudo-likelihood. The inclusion of an extra binomial dispersion parameter in the latter methods did not clearly improve their performance. As is well-documented, biases of these methods in the absence of an extra binomial dispersion parameter are towards the null (Rodriguez and Goldman, 1995) and (Rodriguez and Goldman, 2001), whereas in the presence of extra binomial dispersion biases tended to be less predictable (in either direction). Moreover, we are not convinced about the usefulness of ϕ as a diagnostic tool. The REPL procedure performed poorly in our settings, substantiating the finding reported by Masaoud and Stryhn (2010) that REPL performs mostly as a marginal estimation procedure with no promise for estimation of the variance or an autoregressive parameter. Further research may be needed to assess its accuracy and validity for binary repeated measures data.

Among the marginal procedures, ALR and GEE with either independence or exchangeable correlation at the cluster-level performed similarly and generally well across the range of settings covered. All other attempts to incorporate the additional hierarchical level into the GEE framework produced estimates with serious deficiencies for some of the fixed effects parameters. In situations where the affected parameters are of no interest (or absent), such other schemes may be acceptable but on the other hand show no advantages over the above-mentioned generally acceptable schemes.

Appendix

Table A1: converted (using formula 2.3) marginal estimates from Table 4 obtained by different random effects statistical procedures for two somatic cell count datasets comprising 20 cows in each of 30 herds and measures at $t = 4$ or $t = 8$ time points, respectively. Parameters (see text for model details): β_0 (intercept), β_1 (month), β_2 (parity), β_3 (breed). Coding for estimation procedures with logit link: PQL (2nd order penalized quasi-likelihood), PQLx (2nd order penalized quasi-likelihood with extra binomial dispersion), REPL (restricted pseudo-likelihood), ML (maximum likelihood), MCMC (Bayesian Markov chain Monte Carlo).

t	Par.	Estimation procedures									
		PQL		PQLx		REPL		ML		MCMC	
		Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
4	β_0	-1.239	.193	-1.262	.181	-1.281	.161	-1.319	.190	-1.310	.201
	β_1	.213	.034	.232	.025	.222	.025	.203	.032	.201	.031
	β_2	-.673	.159	-.662	.156	-.606	.128	-.688	.149	-.671	.145
	β_3	.066	.218	.075	.214	.069	.188	.083	.208	.078	.215
8	β_0	-1.080	.182	-1.085	.180	-1.070	.166	-1.089	.179	-1.084	.190
	β_1	.129	.012	.131	.010	.128	.011	.124	.011	.122	.012
	β_2	-.732	.130	-.730	.135	-.682	.114	-.760	.127	-.752	.123
	β_3	-.095	.229	-.092	.228	-.078	.208	-.089	.223	-.073	.238

References

- Alban, L. and Agger, J. F. (1996), "Welfare in Danish dairy herds 1. Disease management routines in 1983 and 1994," *Acta Veterinaria Scandinavica* 37, 49–63.
- Barbosa, B. and Goldstein, H. (2000), "Discrete multilevel response models," *Quality and Quantity*, 34, 323–330.
- Breslow, N. E. and Clayton, D.G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.
- Breslow, N.E. (2003), "Whither PQL?," University of Washington Biostatistics Working Paper Series, 192.
- Browne, W. J. and Draper, D. (2006), "A comparison of Bayesian and likelihood-based methods for fitting multilevel models," *Bayesian Analysis*, 3, 473–514.
- Carey, V., Zeger, S.L. and Diggle, P. (1993), "Modeling multivariate binary data with alternating logistic regressions," *Biometrika*, 80, 517–526.
- Congdon, P. (2003), *Applied Bayesian Modelling*, Wiley, New York.
- De Vliegher, S., Barkema, H. W., Stryhn, H., Opsomer, G., de Kruif, A. (2004), "Impact in dairy heifers of early lactation somatic cell count on somatic cell counts over the first lactation," *Journal of Dairy Science*, 87, 3672–3682.
- Diggle P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002), *Analysis of Longitudinal Data (2nd ed.)*, Oxford University Press, Oxford.
- Dohoo, I. R., Martin, S. W. and Stryhn, H. (2009), *Veterinary Epidemiologic Research, (2nd Edition)*, AVC Inc., Charlottetown, Canada.
- Dohoo, I. R., Meek A. (1982), "Somatic cell counts in bovine milk," *Canadian Veterinary Journal*, 23, 119–125.
- Dohoo, I. R. and Stryhn, H. (2006), "Simulation studies on the effects of clustering," *XIth International Conference of Veterinary Epidemiology and Economics*, Cairns, Australia.
- Evans, B. A., Feng, Z. and Peterson, A. V. (2001), "A comparison of generalized linear mixed model procedures with estimating equations for variance and covariance parameter estimation in longitudinal studies and group randomized trials," *Statistics in Medicine*, 20, 3353–3373.
- Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models (Comments on article by Browne and Draper)," *Bayesian Analysis* 3, 515–534.
- Goldstein, H. (1991), "Nonlinear multilevel models with an application to discrete response data," *Biometrika*, 78, 45–51.
- Goldstein, H. and Rasbash, J. (1996), "Improved approximations for multilevel models with binary responses," *Journal of the Royal Statistical Society, Series A*, 159, 505–513.
- Goldstein, H., Browne, W. J. and Rasbash, J. (2002), "Partitioning variation in multilevel models," *Understanding Statistics*, 1, 223–231.
- Green, M. J., Burton, P. R., Green, L. E., Schukken, Y. H., Bradley, A. J., Peeler, E. J. and Medley, G. F. (2004), "The use of Markov Chain Monte Carlo for analysis of correlated binary data:

- Patterns of somatic cells in milk and the risk of clinical mastitis in dairy cows,” *Preventive Veterinary Medicine*, 64, 157–174.
- Heagerty, P. J. and Zeger, S. L. (2000), “Marginalized multilevel models and likelihood inference,” *Statistical Science*, 15, 1–19.
- Heo, M. and Leon, A. C. (2005), “Comparison of statistical methods for the analysis of clustered binary observations,” *Statistics in Medicine*, 24, 911–923.
- Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R. and Jones, D. R. (2005), “How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS,” *Statistics in Medicine*, 24, 2401–2428.
- Liang, K. Y. and Zeger, S. L. (1986), “Longitudinal Data-Analysis Using Generalized Linear-Models,” *Biometrika*, 73, 13–22.
- Kunthel Bya, Bahjat F. Qaqisha, John S. Preisser, Jamie Perinb, Richard C. Zinkc. (2014), “ORTH: R and SAS software for regression models of correlated binary data based on orthogonalized residuals and alternating logistic regressions,” *Computer Methods and Programs in Biomedicine*, 113, 557–568
- Masaoud, E. A. M. (2009), “Statistical models for binary repeated measures and hierarchical data in veterinary science,” PhD Thesis, Department of Health Management, Atlantic Veterinary College, Charlottetown, Canada.
- Masaoud, E. and Stryhn, H. (2010), “A simulation study to assess statistical methods for binary repeated measures data,” *Preventive Veterinary Medicine*, 93, 81–97.
- McCulloch, C. E. and Searle, S. R. (2008), *Generalized Linear and Mixed Models*, (2nd Edition). Wiley, New York.
- Molenberghs, G. and Verbeke, G. (2005), *Models for Discrete Longitudinal Data*. Springer, New York.
- Moerbeek, M. (2004), “The Consequence of Ignoring a Level of Nesting in Multilevel Analysis,” *Multivariate Behavioral Research*, 39:1, 129–149
- Neuhaus, J. M. (1992), “Statistical methods for longitudinal and clustered design with binary responses,” *Statistical Methods in Medical Research*, 1, 249–273.
- Preisser, J. S., Arcury, T. A. and Quandt, S. A. (2003), “Detecting patterns of occupational illness clustering with alternating logistic regressions applied to longitudinal data,” *American Journal of Epidemiology*, 158, 495–501.
- Olde Riekerink, R. R. G. M., Barkema, H. W. and Stryhn, H. (2007), “The effect of season on somatic cell count and the incidence of clinical mastitis,” *Journal of Dairy Science*, 90, 1704–1715.
- Rabe-Hasketh, S., Skrondal, A. and Pickles, A. (2002), “Reliable estimation of generalised linear mixed models using adaptive quadrature,” *The Stata Journal*, 2, 1–21.
- Rodriguez G, Goldman N. (1995), “An assessment of estimation procedures for multilevel models with binary responses,” *Journal of the Royal Statistical Society, Series A*, 158, 73–89.
- Rodriguez G, Goldman N. (2001), “Improved estimation procedures for multilevel models with

- binary response: a case-study,” *Journal of the Royal Statistical Society, Series A*, 164, 339–355.
- Schukken, Y. H., Grohn, Y. T., McDermott, B. and McDermott, J. J. (2003), “Analysis of correlated discrete observations: background, examples and solutions,” *Preventive Veterinary Medicine*, 59, 223–40.
- Skrondal, A. and Rabe-Hesketh, S. (2007), “Redundant overdispersion parameters in multilevel models for categorical responses,” *Journal of educational and behavioral statistics*, 32, 419–430.
- Snijders, T. A. B. and Bosker, R. J. (2012), *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling (2nd edition.)*. Sage Publishers, London.
- Stryhn, H., Dohoo, I. R., Tillard, E. and Hagedorn-Olsen, T. (2000), “Simulation as a tool of validation in hierarchical generalised linear models,” IXth International Conference of Veterinary Epidemiology and Economics, Breckenridge, Colorado.
- Sturtz, S., Ligges, U. and Gelman, A. (2005), “R2WinBUGS: A package for running WinBUGS from R,” *Journal of Statistical Software*, 12, 1–16.
- Teerenstra, S., Lu, B., Preisser, JS., van Achterberg, T., Borm, GF. (2010), “Sample Size Considerations for GEE Analyses of Three-Level Cluster Randomized Trials,” *Biometrics*, 66, 1230–1237.
- Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S (4th ed.)*. Springer, New York.
- Virginie, R., Kiros, B. and Duncan, C. T. (2005), “A three-level model for binary time-series data: the effects of air pollution on school absences in the Southern California Children’s Health Study,” *Statistics in Medicine*, 24, 1103–1115.
- Wolfinger, R. and O’Connell, M. (1993), “Generalized linear mixed models: a pseudo-likelihood approach,” *Communications in Statistics: Simulation and Computation*, 48, 233–243.
- Yang, M., Goldstein, H. and Heath, A. (2000), “Multilevel models for repeated binary outcomes: attitudes and voting over the electoral cycle,” *Journal of the Royal Statistical Society, Series A*, 163, 49–62.
- Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988), “Models for longitudinal data: a generalized estimating equation approach,” *Biometrics*, 44, 1049–1060.
- Ziegler, A., Kastner, C. and Blettner, M. (1998), “The generalized estimating equations: An annotated bibliography,” *Biometrical Journal*, 40, 115–139.

Received: October 24, 2019

Accepted: May 9, 2020