Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2023;9(2):257-269

Original Article

Addressing Heteroscedasticity in Correlated Binary Data: A Bayesian Mixed Effects Location Scale Approach

Parisa Rezanejad-Asl[1,2], Farid Zayeri[3*], Abbas Hajifathali[4]

[1]Department of Biostatistics and Epidemiology, School of Health, Alborz University of Medical Sciences, Karaj, Iran.
[2]Department of Biostatistics, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[3]Proteomics Research Center and Department of Biostatistics, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[4]Hematopoietic Stem Cell Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

ARTICLE INFO

Received 17.02.2023
Revised 11.03.2023
Accepted 28.03.2023
Published 15.06.2023

Key words:
Correlated binary data;
Heteroscedasticity;
Bayesian mixed-effects model;
Variance modelling;
Random scale effects;
Longitudinal data analysis

ABSTRACT

Introduction: The mixed effects logistic regression model is a common model for analysing correlated binary data as longitudinal data. The between and within subject variances are typically considered to be homogeneous but longitudinal data often show heterogeneity in these variances. This study proposes a Bayesian mixed effects location scale model to accommodate heteroscedasticity in binary data analysis.
Methods: This study was carried out in two stages; first, the simulation study was used to evaluate the accuracy of the proposed model with the Bayesian approach and then the proposed model was applied to a real data. In simulation study, the data were generated from the mixed effects location scale model with different correlations between the random location effect and random scale effect and different sample sizes. In order to evaluate the accuracy of the estimations, the Root Mean Square Error, bias and Coverage Probability were calculated and the deviance information criterion was used to select the appropriate model. At the end we utilized this model to analyse uric acid levels of patients with haematological disorders.
Results: The simulation results show the accuracy of model parameter estimates as well as the correlation between random location and scale effects. They also display that if a random scale effect is present in the data, it should be accounted for in model. Otherwise, the model is forced to assign the within subject variation due to these subject random effects to the error term. The results of real data are also in line with this. The odds of having normal UA levels increases by a factor of 26% per week. Due to the positive value of the covariance parameter, patients with higher mean of UA levels show higher variation in UA levels. Furthermore, the significance of the covariates in the between subject and within subject variances model, as well as the significance of the random scale variance determines the heterogeneity across subjects.
Conclusion: Bayesian mixed effects location scale model provides a useful tool for analysing correlated binary data with heteroscedasticity because it considers data correlation and modelling mean and variance simultaneously. Furthermore, it improves the accuracy of statistical inference in longitudinal studies compared to classic mixed effects models.

*.Corresponding Author: fzayeri@gmail.com

Introduction

Longitudinal data with binary response variables mostly occur in medical investigations, especially in clinical trials. In binary longitudinal data, the recurrence of outcome is a subject that should be considered and the relationship between covariates and the multiple incidents of the outcome is of interest. Analysis of such data is a challenging issue as some individuals are more prone to recurrences than others, and, therefore, the response variable shows positively correlated repeated measures.[1]

There are several approaches for the analysis of correlated longitudinal categorical data. Marginal modeling as one of these approaches provides inferences for parameters averaged over the whole population. The usual method for parameter estimation in marginal models is the generalized estimating equations (GEE).[2,3] Using random effects modeling as another approach provides inferences about the variability between respondents.[4,5] Markov (transition) models are other approaches to evaluate the reasons for the change of the responses. The maximum likelihood method is often used to estimate the parameters of both the random effects and Markov models.[6,7] Tang et al. used Binary logistic regression to analyze China Health and Retirement Longitudinal Study data sets in order to explore the association of midday napping with hypertension, and the 3-step method was used to test the mediation effect of BMI. They concluded that BMI serves as a mediator and that midday naps increase the risk of hypertension.[8] Iddris et al evaluated the effect of gender on blood pressure (BP) over the three BP measurements adjusting for other risk factors of BP in Ghana by using the logistic mixed effects model.[9]

The mixed effects logistic regression models, as an extension of generalized linear models) GLM), are the most common statistical tools for analyzing binary response longitudinal data.[10,11] The observations for the same individuals are correlated at different times and these models consider these correlations by including one or several random effects.[4] The between-subject (BS) and within-subject (WS) variances which refer to random effects and error variances, respectively, are usually assumed to be homogeneous across subjects. However, sometimes this assumption is not supported by the data.[12] Examples of error variance which is systematically related to the explanatory variables are available. Carroll and Ruppert defined the idea of modeling the error variance in terms of explanatory variables.[13] Balazs, Hidegkuti, and De Boeck tried to evaluate participant heterogeneity in item-response data using a logistic regression model in which heterogeneity emerged as a latent random term added to the main effects and covariate dependent terms.[14]

Hedeker et al. introduced the mixed effects location scale (LS) model in 2006; this model is a useful approach for joint modeling of mean and variance structure. They tried to include the covariates in both WS and BS variances to account the heteroscedasticity and model their influences on both variation sources. Moreover, to capture heterogeneity in random errors, they included a random term at subject level into the WS variance modeling. The random location and scale effects can characterize the subject's influence on both mean (location) and variability (scale) of the longitudinal outcome; the random location and

scale effects are correlated to some extent.[15]

The aforesaid modeling heteroscedasticity approaches have been also evaluated in the Bayesian framework. Hoff and Niu introduced a method to parametrize the covariance matrix of a multivariate response vector as a parsimonious quadratic function of explanatory variables. They used the EM-algorithm and MCMC approximation via Gibbs sampling to explain and clarify parameter estimation.[16] Rast et al. modeled the individual differences in level by Bayesian approach and modified using the mixed-effects location scale model proposed by Hedeker et al.[17] Efficient estimation of the regression parameters can be the result of modeling heteroscedasticity; it can cause more precise predictive inferences for some units and less for others compared to models assuming variance homogeneity that yield the same accuracy for all observations.[12]

In this paper, we aimed to generalize the approach developed by Hedeker et al. to the binary outcomes by including the random effects at both the location and scale levels within a Bayesian framework.

The article is organized as follows: Section 2 describes the notation and the model. Section 3 provides a brief overview of Bayesian estimation and the model fit criteria. The analysis of real data is presented in Section 4 and Section 5 shows the results of a simulation study. Finally, the results are discussed in Section 6.

Methods

Model description

Let Y_{ij} denotes the binary outcome (taking values of 0 or 1) for subject i (i=1,…,N) at time j (j=1,…,n_i). A common logistic mixed effects model can be written as:

$$logit\left(P_{ij}\right)=\log\left(\frac{P_{ij}}{1-P_{ij}}\right)=X_{ij}^{'}\beta+Z_{ij}^{'}u_i.$$

$$P_{ij}=\Pr(Y_{ij}=1)=\exp(X_{ij}'\beta+Z_{ij}'u_i)/(1+\exp(X_{ij}'\beta+Z_{ij}'u_i)), \quad (1)$$

where X_{ij} is a p × 1 vector of predictor variables, Z_{ij} is a subset of X_{ij}, β is a vector of fixed effects parameters corresponding to the predictor variables. The random subject effect $u_i \sim N(0,\sigma_{u_i}^2)$ is used to account for the correlation between the repeated measurements on the same subject. In this model, σ_u^2 denotes the between-subjects (BS) variance. Now we add the scaling terms to this framework and introduce the mixed effects location-scale model for binary response data as.[18]

$$logit\left(P_{ij}\right)=\log\left(\frac{P_{ij}}{1-P_{ij}}\right)=\frac{X_{ij}^{'}\beta+Z_{ij}^{'}u_i}{\sigma_{\varepsilon_{ij}}}. \quad (2)$$

We also used a log-linear representation to relate the covariates to the BS and WS variances, as explained in the context of heteroscedastic (fixed-effects) regression models,[19,20] namely:

$$\sigma_{u_i}^2=\exp(v_i'\tau), \quad (3)$$

$$\sigma_{\varepsilon_{ij}}^2=\exp(w_{ij}'\gamma), \quad (4)$$

where τ and γ are fixed-effects parameters which express the degree of influence of their corresponding covariates, v_i and W_{ij}, on the BS and WS variances, respectively. If $\tau = \gamma = 0$, the classic logistic regression with random intercept is achieved. The BS variance is modeled only with subject-varying covariates (such as demographic characteristics like sex,

…) but the WS variance is modeled with both subject-varying and time-varying covariates (such as time). Since the exponential function guarantees a multiplicative factor for certain values of and , the resulting variance will definitely be positive.

To consider the individuals' heterogeneity, we can further extend the model by including random subject effects for a subject `s measurement error (i.e., random scale effects):

$$\sigma^2_{\varepsilon_{ij}} = \exp(w'_{ij}\gamma + \omega_i). \qquad (5)$$

where the random subject (scale) effects ω_i are distributed in the population of subjects with mean 0 and variance σ^2_ω. If the distribution of ω_i is normal, then the WS variance,

$$\log(\sigma^2_{\varepsilon_{ij}}) = w'_{ij}\gamma + \omega_i$$

follows a log-normal distribution. A log-normal distribution is a proper choice for representing variances because of its features such as skewed and nonnegative nature.[21-23] Model[5] allows the WS variance to vary across subjects, beyond the effect of covariates. In this model, u_i is a random effect that influences the location or mean of the individual`s outcome and is a random scale effect that influences an individual`s variance. Thus, the model with both types of random effects is called as mixed effects location scale model. These two random effects are correlated with correlation parameter ρuω, which indicates the degree of association between the random location and scale effects.[15]

Bayesian inferences

The parameter estimation is usually challenging in the mixed effects location scale model for binary data due to computational complexity and convergence failure. Here, we introduced a fully Bayesian approach to estimate parameters in model (2), (3), and (5). In this context, the Markov chain Monte Carlo (MCMC) procedures enabled us to sample the posterior distribution for each parameter and make inference afterwards. Figure 1 illustrates how Bayesian analysis brings together observed data with prior probabilities and a model to obtain the results.

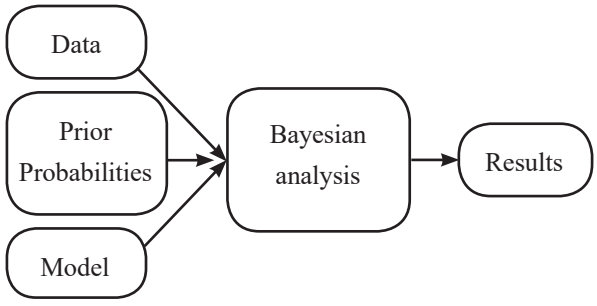

Figure 1. General approaches of the Bayesian methodology

In this study, we utilized vague priors for fixed-effects regression coefficients using the normal distribution, so that, $\beta \sim N(0, 0.001)$, $\tau \sim N(0, 0.001)$ and $\gamma \sim N(0, 0.001)$. Also the diffuse prior was specified for the inverse of the variance of the scale effects using the gamma distribution with small scale and shape parameters, i.e. G(0.001,0.001).

If we suppose that the model parameters, $\Theta = \{\beta, \tau, \gamma\}$, are independent of each other, then the prior density function will be $P(\Theta) = P(\beta)P(\gamma)P(\tau)$. Thus, the joint posterior density of Θ is:

$$P(\Theta \mid Y_{i1},...,Y_{in_i}, X_i) \propto L(\Theta \mid Y_{i1},...,Y_{in_i}, X_i)P(\Theta),$$

where $L(\Theta \mid Y_{i1},...Y_{in_i})$ is the likelihood function and

$$L(\Theta \mid Y_{i1},...Y_{in_i}) = \prod_{i=1}^{N} \int_{u,\omega} \prod_{j=1}^{n_i} P(Y_{ij} = y_{ij} \mid u_i, \omega_i; \beta) f(u_i, \omega_i) \partial u \partial \omega, \quad (6)$$

$$P(Y_{ij} = y_{ij} \mid u_i, \omega_i; \beta) = P_{ij}^{y_{ij}} (1 - P_{ij})^{1-y_{ij}},$$

$$f(u_i, \omega_i) = \frac{1}{(2\pi)\sigma_{u_i}\sigma_{\omega_i}\sqrt{1-\rho_{u\omega}^2}} \exp\{-\frac{1}{2(1-\rho_{u\omega}^2)}$$

$$[(\frac{u_i}{\sigma_{u_i}})^2 - 2\rho_{u\omega}(\frac{u_i\omega_i}{\sigma_{u_i}\sigma_{\omega_i}}) + (\frac{\omega_i}{\sigma_{\omega_i}})^2]\}.$$

In general, the integrals in[6] do not have closed form. Since the numerical approximation of the integrals may be inaccurate, it is not recommended to directly calculate the posterior distribution from the observed data. An alternative approach is to use the MCMC procedure for sampling from posterior distributions, based on,[6] using the Gibbs sampler along with the Metropolis-Hastings algorithm.

We implemented the proposed model using WinBUGS 1.4 software[24] and R2WinBUGS package[25] in R 4.3.0 software. We run the model for 50,000 iterations with 20,000 iterations for burn-in followed by 30,000 samples for estimates using two parallel chains.

Model selection

Here, the deviance information criterion (DIC) was applied as a criterion for choosing the best model.[26] DIC is a Bayesian version of Akaike's information criterion (AIC) which has been commonly used for Bayesian model selection, especially where the posterior distributions of the models is obtained using the MCMC approach.[27] For data y and model parameters Θ, the expression for DIC is as follows

$$DIC = 2\overline{D}(y,\Theta) - D(y,\overline{\Theta})$$

where $\overline{D}(y,\Theta) = E(D(y,\Theta) \mid y)$ is the posterior mean deviance and $D(y,\overline{\Theta})$ is the deviance at the posterior mean of Θ, denoted by $\overline{\Theta}$ (statistical deviance is defined as $D(y.\Theta)=-2\ln(f(y.\Theta)))$. The models with smaller values of DIC are preferred.

Results

Real data application

To illustrate the application of the mixed effects location scale model for binary responses, we used the data from a study on 166 patients with hematological disorders who were admitted for allogeneic transplantation at Taleghani hospital, Tehran, Iran, between 2008 and 2018. The mean±SD age of the patients was 32.09±10.59 (ranged 7 to 57 years) and patients 82(49.4%) were female.

Allogeneic hematopoietic stem cell transplantation (allo-HSCT) rate is increasing nowadays throughout the world; every year there is about 50,000 – 60,000 transplantation.[28] Graft versus host disease (GvHD) after HSCT is an important complication and thus it is a therapeutic challenge.[29] Patients undergoing allo-HSCT showed GVHD prevalence as 20-60%.[21] Other studies showed that the HSCT process changes serum uric acid (UA) levels in allo-HSCT.[31] Also, they described the impact of UA levels (as a sensitive biomarker) on the incidence of GVHD and overall survival in allo-HSCT patients.[32,33]

Here, our aim is to evaluate the trend of UA levels and to determine whether there is a difference between male and female patients in UA levels, adjusting for the patients' age. Serum UA levels were measured at one and two weeks before allo-HSCT, the day of allo-HSCT, one and finally two weeks after allo-HSCT. For males, the normal range for uric acid biomarker is between 3.6 and 8.2, while this range is between 2.3 and 6.0 in females

(based on the determined normal range by the local tab). If UA level was in the normal range, then $y_{ij}=1$, otherwise $y_{ij}=0$.

To begin, we fitted a random-intercept logistic model(1) with vague priors (densities with high spread) using normal distribution, $\beta \sim N(0,0.001)$ and Gamma distribution, $\sigma_u^2 \sim G(0.001,0.001)$, for the inverse of random effect variance. In this model, we did not include any covariates for the variances (i.e. for V_i or W_{ij}). The patients' sex (female=0 and male=1) and time of measurement were considered as the covariates in the final model (adjusting for age). It should also be noted that the interaction between sex and time was not included in the final model because it had not significant effect.

$$logit((y_{ij}=1\,|\,x_{ij}))=logit(P_{ij})=\log(\frac{P_{ij}}{1-P_{ij}})=$$
$$\beta_0+\beta_1*time_j+\beta_2*sex_i+\beta_3*age_i+u_i,$$

Next, to fit the mixed effects location scale, we added these covariates into the log linear models of the BS and WS variances (the time variable was not included in the BS variance model). We used the vague priors for fixed effects regression coefficients using the normal distribution, $\beta \sim N(0,0.001)$, $\gamma \sim N(0,0.001)$ and $\tau \sim N(0,0.001)$ as well as the Gamma distribution with very small scale and shape parameters for the inverse of the variance of the scale effect, $\sigma_\omega^2 \sim G(0.001,0.001)$.

$$logit((y_{ij}=1\,|\,x_{ij}))=logit(P_{ij})=\log(\frac{P_{ij}}{1-P_{ij}})=$$
$$\beta_0+\beta_1*time_j+\beta_2*sex_i+\beta_3*age_i+u_i,$$

$$\sigma_{u_i}^2=\exp(\tau_0+\tau_1*sex_i)$$
$$\sigma_{\varepsilon_{ij}}^2=\exp(\gamma_0+\gamma_1*time_j+\gamma_2*sex_i+\omega_i)$$

Table 1 shows the obtained results from fitting the described models. As can be seen, in the mixed effects model, the tendency to normal UA levels increases significantly from two weeks before allo-HSCT to two weeks after allo-HSCT(odds ratio (OR)=1.27). In other words, the odds of normal UA levels weekly increases by a factor of 27% in these patients. In addition, the estimated OR=0.51 for the sex variable means that odds of having normal UA levels in male patients was about half of the same odds in females.

According to the results of the mixed effects location scale model, in terms of the location modeling, the estimated odds of normal UA levels weekly increases by a factor of 26% in these patients. Also, the estimated OR=0.52 for the sex covariate shows that odds of having normal UA levels in male patients was about half of the same odds in females. For modeling BS variance, the 'sex' effect is significant. Female has less BS variation than male because the sign for the coefficient estimates of 'sex' is positive. The WS variance was modeled via a log link function. Time $(\gamma_1=-0.24)$ and sex $(\gamma_2=0.93)$ were significant predictors of WS variability. The signs of the estimates showed that by increasing the time, the UA levels less varied. The random WS variance (the BS variance of scale) and covariance parameters are both highly significant. The significant variance of the random WS variance effect σ_ω^2 indicates that there was considerable heterogeneity among the patients in terms of their UA levels variation. The covariance parameter $\sigma_{u\omega}$ is estimated to be positive. Thus, patients with higher mean of UA levels exhibit greater variation in UA levels. In other words, these results show that the BS and WS variances in these data are not homogeneous and by considering this heterogeneity, the

Table 1. Mixed effects and the mixed effects location scale model with random intercept

Sub-models	Parameter	Mixed effects model				Mixed effects location scale model			
		Estimate*	OR*	S.D***	CI****	Estimate	OR	S.D	CI
Location	Intercept	-0.76	0.48	0.41	(-1.54,-0.03)	-0.74	0.47	0.35	(-1.45, -0.07)
	Time	0.24	1.27	0.06	(0.10,0.34)	0.23	1.26	0.06	(0.12, 0.34)
	Sex (M/F)	-0.69	0.51	0.23	(-1.1, -0.30)	-0.65	0.52	0.22	(-1.09, -0.21)
WS[a] variance	Intercept					-1.697		0.46	(-2.58, -0.80)
	Time					-0.242		0.06	(-0.37,-0.11)
	Sex (M/F)					0.9252		0.25	(0.45,1.42)
BS[b] variance	Intercept					-2.459		1.07	(-4.08,-0.44)
	Sex (M/F)					2.434		0.49	(1.64,3.98)
Scale	σ_{ω}^2					0.6534		0.65	(0.46, 1.04)
	$\sigma_{u\omega}$					0.9746		0.02	(0.92, 0.99)
DIC		935				803			

*Estimated posterior mean,
**Odds ratio,
***Standard deviation,
****95% equal-tail credible interval
[a] Within Subject
[b] Between Subject

estimation of parameters will be more accurate.

Simulation

To evaluate the performance of the proposed Bayesian approach, we compared the estimators of the mixed effects location scale models with different sample sizes and different correlations between random location effect and random scale effect in terms of their bias, the root mean squared error(RMSE), and the coverage of the 95% highest density interval using MCMC method.

In this context, 1000 longitudinal binary datasets were simulated from the proposed mixed effects location scale model for binary outcomes (2),(3), and (5) regarding four different sample sizes (*N*=50,N=100, *N*=300, *N*=500) each with five time points (*ni*=5) and using the following model parameters and covariates:

- X_1 is a normally distributed covariate with mean 0.5 and variance 0.04. X_2 is a binary variable taking value 1 with p=0.5 (which is generated from a Bernoulli distribution with success probability p=0.5)
- The true values for the fixed effect are set as:
- β= (-0.64,0.22,0.06)', γ= (0.1,0.07, -0.01)', τ= (-2.13,0.59)'.
- The correlation terms of random location and random scale effects are set to three different values ($\rho_{u\omega}$=0, $\rho_{u\omega}$=0.5, $\rho_{u\omega}$=0.9).

The BS and WS variances were permitted to vary between and within subjects, respectively. The BS variance was modeled by log-linear models with X_1 and the WS variance was modeled by log-linear models with X_1 and X_2 including a random subject scale (ω_i) parameter to define the variability in the WS variance which

is not explicated by the covariates. We used these simulated datasets to fit similar models (2),(3),(5) and the MCMC sampling schema that is used for the real data analysis. Table 2, 3 and 4 shows the obtained results.

As can be seen in the results, the estimated parameters of the proposed model and the scale parameters are nearly unbiased with small RMSEs (an estimate of the model's ability to predict the target value (accuracy)). Also, the coverage (the probability that a highest density interval will include the true value of interest) of the 95% highest density interval for all of the parameters was above 90% in different scenarios. The suggested model provided unbiased estimates of the true values of the correlation terms of random location and random scale effects in different

Table 2. Location scale mixed effects model with random intercept based on simulation with $\rho_{u\omega}$=0.

Sub model	Parameter	True value	N=50			N=100			N=300			N=500		
			Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE
Location	β_0	-0.64	-0.100	0.89	0.125	-0.096	0.91	0.115	-0.068	0.93	0.092	-0.041	0.93	0.083
	β_1	0.22	0.023	0.91	0.098	0.021	0.92	0.078	0.017	0.90	0.051	0.018	0.92	0.048
	β_2	0.06	0.021	0.94	0.191	0.021	0.94	0.136	0.019	0.96	0.100	0.016	0.97	0.081
Scale	w_0	0.1	0.034	0.91	0.145	0.032	0.91	0.107	0.031	0.93	0.081	0.029	0.94	0.078
	w_1	0.07	0.028	0.88	0.074	0.024	0.89	0.094	0.015	0.91	0.057	0.013	0.92	0.025
	w_2	-0.01	-0.022	0.90	0.188	-0.017	0.94	0.158	0.014	0.93	0.085	0.011	0.97	0.009
	ν_0	-2.13	-0.093	0.90	0.219	-0.089	0.89	0.182	-0.080	0.91	0.102	-0.057	0.93	0.078
	ν_1	0.59	0.178	0.89	0.227	0.172	0.90	0.192	0.153	0.92	0.187	0.125	0.90	0.107
	σ^2_ω	0.53	0.113	0.88	0.271	0.109	0.89	0.243	0.099	0.90	0.102	0.096	0.91	0.095
	$\rho_{u\omega}$	0	0.0086	0.92	0.154	0.008	0.93	0.147	0.0067	0.96	0.135	0.006	0.96	0.128

Table 3. Location scale mixed effects model with random intercept based on simulation with $\rho_{u\omega}$=0.5.

Sub model	Parameter	True value	N=50			N=100			N=300			N=500		
			Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE
Location	β_0	-0.64	-0.102	0.9	0.119	-0.094	0.92	0.104	-0.054	0.92	0.083	-0.041	0.94	0.081
	β_1	0.22	0.021	0.91	0.093	0.019	0.91	0.062	0.017	0.93	0.049	0.017	0.93	0.051
	β_2	0.06	0.028	0.94	0.188	0.028	0.95	0.136	0.022	0.97	0.097	0.019	0.99	0.074
Scale	w_0	0.1	0.041	0.92	0.142	0.037	0.94	0.117	0.034	0.91	0.107	0.031	0.93	0.077
	w_1	0.07	0.027	0.89	0.078	0.021	0.91	0.081	0.017	0.93	0.052	0.012	0.95	0.021
	w_2	-0.01	-0.020	0.91	0.183	-0.017	0.95	0.143	-0.013	0.95	0.069	-0.009	0.97	0.009
	ν_0	-2.13	-0.093	0.90	0.210	-0.088	0.91	0.155	-0.082	0.90	0.128	-0.076	0.91	0.083
	ν_1	0.59	0.177	0.88	0.221	0.171	0.90	0.223	0.152	0.91	0.206	0.112	0.92	0.099
	σ^2_ω	0.53	0.107	0.92	0.251	0.099	0.93	0.245	0.097	0.92	0.093	0.097	0.94	0.087
	$\rho_{u\omega}$	0.5	0.052	0.96	0.171	0.048	0.96	0.178	0.042	0.94	0.127	0.034	0.95	0.093

Table 4. Location scale mixed effects model with random intercept based on simulation with $\rho_{u\omega}$=0.9.

Sub model	Parameter	True value	N=50			N=100			N=300			N=500		
			Bias	Cover-age	RMSE	Bias	Cover-age	RMSE	Bias	Cover-age	RMSE	Bias	Cover-age	RMSE
Location	β_0	-0.64	-0.101	0.89	0.110	-0.082	0.91	0.089	-0.041	0.91	0.082	-0.039	0.92	0.078
	β_1	0.22	0.019	0.93	0.091	0.014	0.92	0.051	0.018	0.94	0.043	0.015	0.93	0.039
	β_2	0.06	0.022	0.94	0.184	0.019	0.96	0.111	0.013	0.98	0.074	0.013	0.99	0.069
Scale	w_0	0.1	0.032	0.90	0.142	0.032	0.91	0.094	0.025	0.92	0.072	0.029	0.91	0.052
	w_1	0.07	0.021	0.92	0.075	0.018	0.94	0.061	0.009	0.96	0.018	0.004	0.96	0.014
	w_2	-0.01	-0.018	0.94	0.181	-0.014	0.97	0.112	0.009	0.97	0.039	-0.007	0.99	0.003
	ν_0	-2.13	-0.081	0.92	0.201	-0.073	0.92	0.125	-0.075	0.91	0.112	-0.035	0.93	0.055
	ν_1	0.59	0.159	0.87	0.214	0.143	0.88	0.191	0.101	0.92	0.147	0.008	0.94	0.091
	σ^2_ω	0.53	0.098	0.91	0.241	0.095	0.93	0.233	0.089	0.95	0.098	0.093	0.94	0.073
	$\rho_{u\omega}$	0.9	0.043	0.95	0.193	0.041	0.94	0.180	0.024	0.94	0.094	0.001	0.96	0.082

Table 5. The percentage of times that the mixed effects location scale model has been chosen over the mixed effects model based on the DIC measure

Sample size	Correlation between random location effect and random scale effect		
	$\rho_{u\omega}$=0	$\rho_{u\omega}$=0.5	$\rho_{u\omega}$=0.9
50	81%	83%	87%
100	84%	91%	89%
300	92%	94%	95%
500	98%	97%	98%

scenarios (three different values for $\rho_{u\omega}$). When the sample size increases, an improvement in the accuracy could be observed. These results can also be seen in Figures 2, 3 and 4. In each scenario, we also used the DIC criterion for choosing the best-fitted model. As shown in Table 5, the DIC of mixed effects location scale models were generally smaller than those of the mixed effects models. On the other hand, based on the results of Table 2,3 and 4, higher correlation between the random location effect and random scale effect leads to more accurate estimation of the parameters (Figure 2, 3 and 4). Additionally, Table 5 shows that as the $\rho_{u\omega}$ increases, the percentage of times that the mixed effects location scale model is chosen over the mixed effects model based on the DIC measure also increases.

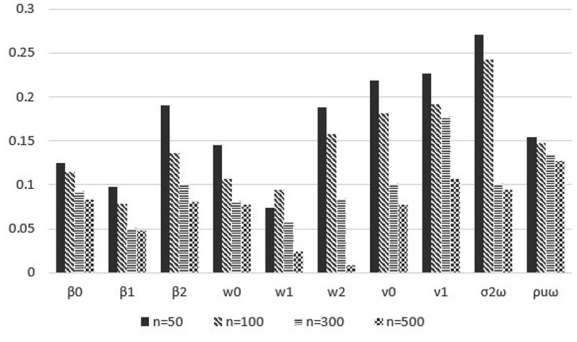

Figure 2. Comparison of mean squared error with different sample size by model parameters for $\rho_{u\omega}$ =0

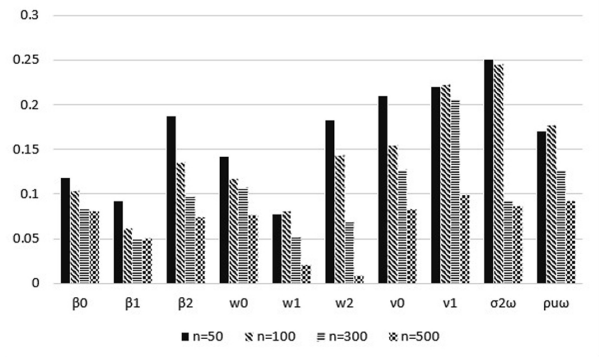

Figure 3. Comparison of mean squared error with different sample size by model parameters for $\rho_{u\omega}$ =0.5

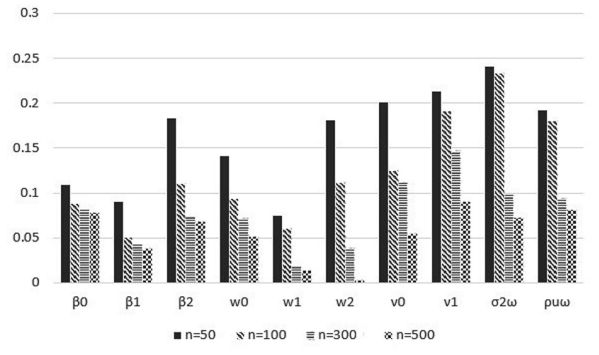

Figure 4. Comparison of mean squared error with different sample size by model parameters for $\rho_{u\omega}$ =0.9

In summary, based on the simulation results, modeling BS and WS variances seems to be important. Therefore, the proposed mixed effects location scale model demonstrates better performance compared to the mixed effects model in the analysis of correlated binary outcomes.

Discussion

In this article, we proposed a mixed effect location scale model for the analysis of binary outcome to deal with heterogeneous longitudinal data within a Bayesian framework. This model holds heterogeneity in BS and WS variances through inclusion of subject- and time-varying covariates. Moreover, this model enables us to explain the heterogeneity in WS variance (which cannot be explained by the covariates) by including a random effect at subject level on WS variance. The random location and scale effects are assumed to be correlated.

Simulation results showed that the proposed mixed effect location scale model is better than the mixed effect because the bias and RMSE for the proposed model are mostly lower than those of the mixed effect model. We also fitted our proposed mixed effect location scale model and mixed effect model on a real dataset of patients` UA levels with hematological disorders who had allogeneic hematopoietic stem cell transplantation. In terms of the mean parameters, sex and time had a significant effect on the UA level. Since Variables such as metabolic syndrome, diabetes, cardiovascular disease, etc., have an impact on changes in UA levels, these confounding variables have been controlled between patients in this study. The odds of having normal UA levels will increase by increasing time. In terms of the WS variance model, time decreased the heterogeneity of a patient`s responses, whereas sex increased this variance. For BS variance, males were more varied in the UA level. The random WS variance (the BS variance of scale) and covariance parameters are both highly significant. We showed that the proposed model was better than the mixed effects model for this data.

Due to computational constraints, the parameter estimation for mixed effects location scale model is challenging. So, a Bayesian approach is proposed to estimate parameters in this model. The main limitation of the research was the lack of longitudinal studies with appropriate follow-up time for including in the

proposed model. In order to accurately estimate the model parameters, it seems necessary to conduct further studies to investigate the outliers and influential observations, especially the variance parameters.

Conclusions

Using mixed effects location scale model for analyzing correlated binary data, in addition to considering data correlation, it deals with modeling mean and variance simultaneously and improves the estimation of model parameters. By considering the heterogeneity of variances across the subjects, this model will have better estimation of the parameters. The Bayesian approach of this model could be a good alternative to the classical approach due to the large number of model parameters (location and scale model parameters) and also because the maximum likelihood function of this model is not closed form.

As a suggestion for future studies, the researchers might use the proposed location-scale mixed effects model in modeling correlated ordinal or count data. In the methods section, we included only a single random term in the location part of the model. This could be generalized to models with multiple location random terms. We also assumed that the random location effects are normally distributed and the random scale effects are log-normally distributed, thus this can be tested using the approach proposed by Liu and Yu for estimating models with non-normal random effects.[34] Another extension of our suggested method could be the simultaneously analysis of multiple binary responses using multivariate mixed effects location scale model.

Conflict of interest

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript. This study has no financial sponsor.

References

1. Glynn, RJ, Rosner, B. Multivariate Methods for Binary Longitudinal Data. Encyclopedia of Biostatistics; 2005.

2. Boateng, EY, Abaye DA. A Review of the Logistic Regression Model with Emphasis on Medical Research. 2019. Journal of Data Analysis and Information Processing.

3. Molenberghs G, Lesaffre E. Marginal modeling of correlated ordinal data using a multivariate plackett distribution. J Am Stat Assoc. 1994;89(426):633–44.

4. Diggle P, Diggle DMSPJ, allgemeine tierzucht F, Press OU, Diggle PJ, Heagerty P, et al. Analysis of Longitudinal Data (2rd ed.). OUP Oxford; 2002. Available from: https://books.google.com/books?id=kKLbyWycRwcC.

5. Verbeke G, Lesaffre E. Linear the Model With Heterogeneity Population in. J Am Stat Assoc. 2012;91(433):217–21.

6. Reuter M, Hennig J, Netter P, Buehner M, Hueppe M. Using Latent Mixed Markov Models for the choice of the best pharmacological treatment. Stat Med. 2004;23(9):1337–49.

7. Chung H, Park YS, Lanza ST.

Latent transition analysis with covariates: Pubertal timing and substance use behaviours in adolescent females. Stat Med. 2005;24(18):2895–910.

8. Tang D, Zhou Y, Long C, Tang S. The Association of Midday Napping With Hypertension Among Chinese Adults Older Than 45 Years: Cross-sectional Study. JMIR Public Health Surveill. 2022 Nov 22;8(11):e38782.

9. Iddrisu, AK., Besing Karadaar, I., Gurah Junior, J. et al. Mixed effects logistic regression analysis of blood pressure among Ghanaians and associated risk factors. Scientific Report. 2023; 13(1).

10. Molenberghs G, Verbeke G. Models for Discrete Longitudinal Data [Internet]. Springer New York; 2006. (Springer Series in Statistics). Available from: https://books.google.com/books?id=aoZ4QljK4YwC

11. Fitzmaurice GM, Laird NM, Ware JH. Applied Longitudinal Analysis [Internet]. Wiley; 2011. (Wiley Series in Probability and Statistics). Available from: https://books.google.com/books?id=qOmxRtdNJpEC

12. Kapur K, Li X, Blood EA, Hedeker D. Bayesian mixed-effects location and scale models for multivariate longitudinal outcomes: An application to ecological momentary assessment data. Stat Med. 2015;34(4):630–51.

13. Carroll RJ, Ruppert D. Transformation and Weighting in Regression [Internet]. Taylor & Francis; 1988. (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). Available from: https://books.google.com/books?id=I5rGEPJd57AC

14. Balázs K, Hidegkuti I, De Boeck P. Detecting heterogeneity in logistic regression models. Appl Psychol Meas. 2006;30(4):322–44.

15. Hedeker D, Mermelstein RJ, Demirtas H. An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. Biometrics. 2008;64(2):627–34.

16. Hoff PD, Niu X. A covariance regression model. Stat Sin. 2012;22(2):729–53.

17. Rast P, Hofer SM, Sparks C. Modeling Individual Differences in Within-Person Variation of Negative and Positive Affect in a Mixed Effects Location Scale Model Using BUGS/JAGS. Multivariate Behav Res. 2012;47(2):177–200.

18. Vahabi N, Kazemnejad A, Datta S. A Marginalized Overdispersed Location Scale Model for Clustered Ordinal Data. Sankhya B. 2018;80:103–34.

19. Aitkin M. Modelling Variance Heterogeneity in Normal Regression Using GLIM. Appl Stat. 1987;36(3):332.

20. White H. Estimating Regression Models with Multiplicative Heteroscedasticity. Econometrica. 1976;48(4):817–38.

21. Vasseur H. Prediction of tropospheric scintillation on satellite links from radiosonde data. IEEE Trans Antennas Propag.

1999;47(2):293–301.

22. Gill N, Hedeker D. Fast estimation of mixed-effects location-scale regression models. Stat Med. 2023 Apr 30;42(9):1430-1444.

23. Renò R, Rizza R. Is volatility lognormal? Evidence from Italian futures. Phys A Stat Mech its Appl. 2003;322:620–8.

24. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual.2003. Available from: internet: http://www.mrc-bsu.cam.ac.uk/bugs.

25. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, & Rubin DB. Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. 2013.

26. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc Ser B Stat Methodol. 2002;64(4):583–616.

27. Meyer, R. Deviance Information Criterion (DIC). Wiley StatsRef: Statistics Reference Online.2016; 1–6.

28. Aljurf M, Weisdorf D, Alfraih F, Szer J, Müller C, Confer D, et al. "Worldwide Network for Blood & Marrow Transplantation (WBMT) special article, challenges facing emerging alternate donor registries." Bone Marrow Transplant [Internet]. 2019;54(8):1179–88.

29. Chao N. Clinical manifestations, diagnosis and grading of acute graft-versus-host disease. UpToDate, Negrin R, Accessed sep 2022. Available from: https://www.uptodate.com/contents/clinical-manifestations-diagnosis-and-grading-of-acute-graft-versus-host-disease.

30. Jagasia M, Arora M, Flowers MED, Chao NJ, McCarthy PL, Cutler CS, et al. Risk factors for acute GVHD and survival after hematopoietic cell transplantation. Blood. 2012;119(1):296–307.

31. Cannell P, Herrmann R. Urate metabolism during bone marrow transplantation. Bone Marrow Transplant. 1992;10:337–9.

32. Haen SP, Eyb V, Mirza N, Naumann A, Peter A, Löffler MW, et al. Uric acid as a novel biomarker for bone-marrow function and incipient hematopoietic reconstitution after aplasia in patients with hematologic malignancies. J Cancer Res Clin Oncol. 2017;143(5):759–71.

33. Ostendorf BN, Blau O, Uharek L, Blau IW, Penack O. Association between low uric acid levels and acute graft-versus-host disease. Ann Hematol. 2015;94(1):139–44.

34. Liu L, Yu Z. A likelihood reformulation method in non-normal random effects models. Stat Med. 2018;27:3105–3124.