https://doi.org/10.4038/sljastats.v24i1.8090

Sri Lankan Journal of Applied Statistics, Vol (24-1)

A Goodness of Fit Test for a Survival and Count Bayesian Joint Model: In the Presence of Clusters

K.U.S. Kumaranathunga* **and M.R. Sooriyarachchi**

Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka.

*Corresponding author: udaraskumaranathunga@gmail.com
https://orcid.org/0009-0007-0019-3913

Received: 8th May 2023/ Revised: 22nd May 2023/ Published: 30th June 2023

ABSTRACT

Bayesian statistical model fitting was an uncommon approach until recently, causing a lack of assessment techniques for these models. However, with the enhancement of computational facilities and advanced estimation techniques, Bayesian models have become popular. Though there are developed goodness of fit (GOF) tests available for classical multilevel models including joint modelling of mixed responses, there is no suitable model based GOF test to be applied on such a model which is fitted under a Bayesian framework. Therefore, this study focused on developing a suitable GOF test for multilevel Bayesian joint models having survival and count responses which are two frequently occurring data types in many fields. The novel test is developed mainly based on four classical GOF tests, including the well-known Hosmer-Lemeshow test and, the Bayesian concepts such as Bayesian credible intervals and regions. In addition, a simulation study has been used to examine the properties of the GOF test together with an application to a real-life example. The novel test performed well in terms of power and acceptable in terms of Type I error rates. Overall, the test performed well with small sample sizes.

Keywords: Bayesian, Markov chain Monte Carlo (MCMC), Goodness of fit (GOF), Mixed response joint model, Simulation, Properties of the test

1 Introduction

The research is conducted with the aim of suggesting a method to assess the goodness of fit (GOF) of a joint Bayesian model of bivariate responses in the presence of cluster variations (i.e., under multilevel modelling). Multilevel

modelling which introduces random effects into the model has become a popular area in statistical modelling due to its applicability in various disciplines. This is widely used in medical, biological, educational and social sciences to overcome the consequences of ignoring hierarchical data structures such as producing underestimated standard errors of regression coefficients (Agresti, 2003). Furthermore, among the two types of statistical modelling techniques namely the classical (frequentist) and the Bayesian, the popularity of Bayesian modelling has been risen in recent decades with the vast growth of computational power (Dunson & Herring, 2005, Bello et al. 2009, Chen et al., 2016). Providing precise estimates for posterior distributions for small samples when using informative priors is a special advantage of Bayesian modelling (Gelman et al., 2004). Thus, the study was focusing mainly on Bayesian modelling in the presence of the hierarchical data structure where the estimation is achieved by a sampling technique, namely the Markov chain Monte Carlo (MCMC) approach. In addition, the basis of this study is built on a joint model of common mixed responses, the survival and the count. Although the joint modelling of these two variables is challenging due to their different distributional nature, it may usually outperform the separate univariate analysis because of the natural correlation among the variables (Sunethra & Sooriyarachchi, 2020, Hapugoda & Sooriyarachchi, 2018).

The literature suggests that there is no model based GOF test to assess adequacy of multilevel joint Bayesian models which are essentially fitted for survival and count responses. Perera, Sooriyarachchi, and Wickramasuriya (2016) have proposed a GOF test for multilevel logistic models where there is a single binary response. More recently, Adhikari and Sooriyarachchi (2020) developed a GOF test for clustered survival and count data based on the tests developed by Hosmer and Lemeshow (1980), Lipsitz et al., (1996) and Perera et al. (2016). All these tests have been developed based on a classical statistical modelling, thus cannot be used to assess the GOF of a joint Bayesian model. Therefore, the main objective of this research is to suggest a GOF test for such models. Starting with the initial idea of developing such a test, the scope of the study was narrowed down to be adapted for small samples since Bayesian analysis works well even for small sample situations. Therefore, the novel development will give rise to the GOF assessment in the Bayesian paradigm which is more practical and better equipped for modelling small sample schemes. The test was developed based on the joint model formulation of Discrete Time Hazard Model for the survival data and Normal model for the log count with Bayesian modelling, by Hapugoda and Sooriyarachchi (2018). Moreover, to assess the performance of the developed test, the two properties, Type I error and power of the test were examined using a simulation study. To generate data for the simulation study, MCMC procedure in SAS software version 9.4 is used. As another objective, the developed test was applied to a real-life dataset to investigate how well the test performs in practice.

The paper is organized into five sections, of which the first section provides an introduction to the study giving some literature related to the research problem and the rest of the paper is organized as follows. Section 2 discusses the methodologies associated with the study and novel GOF test procedure. Section 3 describes the simulation studies to assess the properties of the test. Section 4 gives an application to the developed test. Finally, Section 5 concludes the paper.

2 Theory and Methodology

2.1 Generalized Linear Mixed Models

To deal with the multilevel data structure of the study, generalized linear mixed modelling (GLMM) which is an extension of generalized linear modelling (GLM) that allows cluster-specific components to model is used. Let y_{ij} represents an observation j in cluster i, where j = 1,2, ..., n. The number of observations can vary from cluster to cluster. Let x_{ij} be the vector of explanatory variables for fixed-effect parameters β. Let u_i be the vector of random effects for cluster i, which is the same for all observations within a cluster and z_{ij} represents the covariate vector of random effects. A GLMM extends an ordinary GLM by defining conditional expectation; μ_{ij}=E(Y_{ij} $|u_i$). The GLMM has the general form g(μ_{ij}) = $x_{ij}^{'}\beta + z_{ij}^{'}\mu$ where g(.) is the link function and u_i is assumed to follow a normal distribution with zero mean and a constant variance. (Agresti, 2003).

2.2 Bayesian Approach and Estimation Methods

The parameters in Bayesian are random variables and treated as degrees of belief. As its name implies, the ”Bayesian methods” use the well-known Bayes' theorem as the baseline. Thus, the conclusions about an unknown parameter θ are made in terms of conditional probability statements on the observed values of y: $p(\theta|y)$, where $p(\theta|y) \propto p(\theta)p(y|\theta)$ for a prior distribution $p(\theta)$ and a data distribution $p(y|\theta)$. (Goldstein, 2011). There are two types of prior distributions used in Bayesian modeling, namely the non-informative priors and the informative priors (Browne & Draper, 2006). In the study, non-informative priors are used when fitting the models, as with just a little prior information about the parameters and the data, these can be set to keep the Bayesian estimation in the correct direction (Goldstein, 2011).

2.2.1 Bayesian Hierarchical Models

Due to the hierarchical behaviour of the data in the study, multiple parameters are involved as used in many statistical applications. This implies a need for a joint probability model to reflect the dependence between these parameters (Gelman et al., 2004). Suppose, θ_j be the parameter of interest to be estimated which belongs to j^{th} group and θ has a prior distribution with parameter

vector ϕ (hyper-parameters), for j = 1, 2,…, J. The hierarchical structure occurred for these models, since ϕ has its own prior distribution, p(ϕ) as ϕ is not known. Thus, the appropriate joint prior distribution for the vector (ϕ ,θ) is, p (ϕ,θ)=p(ϕ) p($\theta|\phi$). And the joint posterior distribution is, p (ϕ,θ| y)$\propto$ p (ϕ,θ) p(y |ϕ,θ)=p (ϕ,θ) p(y |θ), since p(y|ϕ,θ) depends only on θ; the y is affected from hyper-parameters only through θ. (Gelman et al., 2004).

2.2.2 Markov chain Monte Carlo Simulations

Markov chain Monte Carlo (MCMC) is a simulation procedure used to fit Bayesian models which draws values of a parameter θ by approximating distributions and correcting the draws to better approximate a target posterior distribution. The samples are drawn sequentially, as depending only on the previous value drawn; hence, the notion of the Markov chain. For converging to a specific target distribution, the approximated distributions are improved at each step in the simulation, using the Markov property (Gelman et al., 2004). Among various algorithms of Markov chain simulations, the Gibbs sampler and the Metropolis-Hastings (MH) algorithm are widely used. However, a recent comparative study on Bayesian MCMC methods in the multilevel context by Karunarasan, Sooriyarachchi, and Pinto (2021) showed that the MH algorithm behaves superior to the Gibbs sampler for small sample sizes, by comparing two-level random intercept models. Concepts associated with MCMC algorithms such as burn-in, thinning and chain length are explained in Karunarasan et al. (2021).

2.2.3 Summarizing Posterior Inference

The Bayesian summary statistics of the posterior probability distribution which are used in this study are Bayesian Credible Intervals and Deviance Information Criterion (DIC). Bayesian Credible Intervals: For this study, a type of Bayesian credible interval, the highest posterior density (HPD) intervals are used to assess the model adequacy, as p-values are not incorporated with Bayesian analysis. The key purpose of using these is to describe and summarise the uncertainty of unknown parameters to be estimated (SAS/STAT 15.1 User's Guide). The 95% HPD interval gives the 95% most credible values. The joint Bayesian model assessment based on 95% credible intervals is discussed in section 2.5. Deviance Information Criterion (DIC): The Deviance Information Criterion (DIC) is used for selecting the best joint model in the application of the study (section 4.2) as the Bayesian model assessment tool. The smaller the DIC, the better the fit of the model.

2.3 Joint Modelling of Survival and Count data

Modelling of Survival and Count variables by taking them as a bivariate response is limited especially where the data being in a hierarchical structure. In such scenarios, it is said that better to model these variables together within

joint modelling using random effects as it accounts for the clustering effect as well as the correlation among responses (Hapugoda & Sooriyarachchi, 2018; Sunethra & Sooriyarachchi, 2020). One such development is the joint survival and count model by combining Discrete Time Hazard Model (DTHM) with the Poisson regression by Hapugoda and Sooriyarachchi (2018). To overcome the difficulties arising due to the nature of the two response variables (i.e., survival variable is continuous while the count is discrete), for modelling the survival time DTHM has been used by converting the survival time into an appropriate discrete time scale. For a more complex method of fitting a joint model of these mixed responses that uses different random effects and complex continuous survival distributions, refer Sunethra & Sooriyarachchi (2020). Moreover, the joint Bayesian model fitted for these responses by Hapugoda and Sooriyarachchi (2018) is explained in the following section.

2.4 The joint Bayesian model used as the basis for this study

The joint model of DTHM with Poisson model, for survival and count responses using Bayesian methods (Hapugoda & Sooriyarachchi, 2018) is used as the basis for the current GOF test development. For simplicity, the model formulation is mentioned here using the same set of notations that will be used in the upcoming sections. As there are two responses, the subscripts 1 and 2 (i = 1, 2) are used to denote them separately. Suppose the notation g indicates the event of interest is happening in time interval g. Let Y_{1gjk} denote the binary (survival) response measured on the jth individual belonging to the kth cluster with a probability of success π_{1gjk}, and let Y_{2jk} denote the count response measured on the jth individual belonging to the kth cluster with an expected count λ_{2jk}. Moreover, X_{jk} denotes the explanatory variables. For $Y = (Y_1, Y_2)$,

$$l_1(Y^{'}_{1gjk}) = \text{logit}(\pi_{1gjk}) = \beta_{01} + v_{ok} + \sum_{g=1}^{m-1} \alpha_g T_{gjk} + \beta_1 X_{jk} \tag{1}$$

and

$$l_2(Y^{'}_{2jk}) = \log(\lambda_{2jk}) = \beta_{02} + v_{ok} + \beta_2 X_{jk} \tag{2}$$

where $v_{ok} \sim N(0, \sigma_v^2)$. Here, l_i denote the GLM link functions for $i = 1, 2$ where l_1 is the logit link and l_2 is the log link. Note: Here the second response variable Y_{2jk} has been directly modeled as the log of count ($\log(Y_{2jk}) = LY_{2jk}$) which is normally distributed with mean μ_{2jk}. Hence,

$$l_2'(LY_{2jk}) = \mu_{2jk} = \beta_{02} + v_{ok} + \beta_2 X_{jk} \tag{3}$$

where l_2' denotes the identity link.

2.5 The steps of novel goodness of fit test of a joint Bayesian model for survival and count responses

The notations used in the succeeding workflow provide the same meaning as when defining the joint Bayesian model for survival and count responses, in section 2.4.

Step 1: The joint Bayesian model is fitted as defined in section 2.4, and the model parameters are estimated using the MCMC method defining appropriate prior distributions.

Step 2: Using the model, for each jth individual nested within kth cluster, the fitted values of the binary survival data and log count data ($\hat{\pi}_{1gjk}$, $\hat{\mu}_{2jk}$) are obtained.

Step 3: First, $\hat{\pi}_{1gjk}$ are sorted in ascending order within each cluster. Then the sorted fitted values are grouped into 5 groups (Balakrishnan & Sooriyarachchi, 2018) within each cluster, and ranks are assigned as $G = 1, 2, \ldots, 5$. The Hosmer and Lemeshow test (1980) approach can be applied within clusters according to Perera et al. (2016).

Step 4: Then to capture the $G = 5$ groups, four indicator variables are introduced as follows.

$$I_{G\_1jk} = \begin{cases} 1, & \text{if } \hat{\pi}_{1gjk} \text{ is in group } G \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $G = 2, 3, 4, 5$ as the '1' is considered as the reference group. Afterwards, the observations are re-arranged as per the observation ID (i.e., the data set is put in order as before the data are sorted). This means the ranks are pooled over the clusters defining the overall 5 groups assigned for the data (Perera et al., 2016).

Step 5: Similarly, the estimated log counts $\hat{\mu}_{2jk}$ are sorted in ascending order within each cluster, and then the sorted values are grouped into 5 groups within each cluster. Subsequently, the ranks $G = 1, 2, \ldots, 5$ are assigned, and four indicator variables are introduced.

$$I_{G\_2jk} = \begin{cases} 1, & \text{if } \hat{\mu}_{2jk} \text{ is in group } G \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $G = 2, 3, 4, 5$ and the first group is taken as the reference group. Then, the data set is re-arranged as per the observation ID.

Step 6: The four indicator variables are included in the joint model of interest,

resulting in the following alternative model. For the binary survival response,

$$\text{logit}(\pi_{1gjk}) = \beta_{01} + v_{ok} + \sum_{g=1}^{m-1} \alpha_g T_{gjk} + \beta_1 X_{jk} + \sum_{G=2}^{5} \gamma_{G\_1} I_{G\_1jk} \qquad (6)$$

and for the log count,

$$\mu_{2jk} = \beta_{02} + v_{ok} + \beta_2 X_{jk} + \sum_{G=2}^{5} \gamma_{G\_2} I_{G\_2jk} \qquad (7)$$

The indicator variables introduce new parameters ($\gamma_{2\_1}, \gamma_{3\_1}, \ldots, \gamma_{4\_2}, \gamma_{5\_2}$) into the joint model. The revised model is estimated using the MCMC method.

Step 7: If all these parameters are simultaneously equal to zero, then the joint Bayesian model is correctly specified. Hence, the hypotheses to be tested are:

$$H_0 : \gamma_{2\_1} = \gamma_{3\_1} = \gamma_{4\_1} = \gamma_{5\_1} = \gamma_{2\_2} = \gamma_{3\_2} = \gamma_{4\_2} = \gamma_{5\_2} = 0$$

H_1 : At least one γ_{G_i} (for $G = 2, \ldots, 5$ and for $i = 1, 2$) is not equal to zero.

Step 8: Bayesian 95% credible intervals for the 4 indicators in each marginal model are assessed to test the above hypotheses, at an overall $\alpha = 0.05 \times 4 = 0.2$ with respect to the credible region $= 0.8$ (This is further explained in section 3.3). If zero is included in each credible interval that belongs to the 4 indicator variables, we do not reject H_0, implying that all the parameters are simultaneously equal to zero. Therefore, the fitted joint Bayesian model is adequate. If zero is not included in at least one credible interval, we reject H_0, implying that at least one of the parameters is not equal to zero. Therefore, the fitted joint Bayesian model is not adequate.

3 Simulation Study

The objective of the simulation study is to assess Type I error and power of the developed test. The data for the simulation study are generated using SAS macros based on the MCMC Procedure in SAS. Different parameters and combinations are considered, and for each scenario, 1000 datasets are generated. The two properties are assessed using the parameter estimates and their credible intervals, which are derived from these datasets (Goldstein, 2011). The design of the simulation procedure and the model fitting is similar for each of the combinations used in the study.

3.1 Parameters and Combinations used for the Simulations

Let the binary (survival) response be denoted as Y_{1gjk} with the probability of success π_{1gjk}. The log count is denoted as LY_{2jk} with the expected log count μ_{2jk}, where the count response is denoted as Y_{2jk} with the expected count λ_{2jk}. Note that these subscripts 1 and 2 are used for the survival model and the log count model, respectively, and j and k denote the second and third levels. The common explanatory variable is represented by X_{jk}, and the survival time indicators in the survival model are represented by T_1 and T_2.

To simulate binary (Y_{1gjk}) and Poisson (Y_{2jk}) data before fitting the joint model, the model coefficients were set either through a trial-and-error method or by directly using values obtained in similar studies (Perera et al., 2016; Adhikari and Sooriyarachchi, 2020). The fixed effect coefficients were chosen as $\beta_{01} = 0.5694$, $\beta_{02} = -0.25$, $\beta_1 = -0.725$, $\beta_2 = -0.2$, $\alpha_1 = 0.02$, and $\alpha_2 = 0.025$, where β_{01} and β_{02} represent the two intercepts, β_1 and β_2 are the slope parameters, and α_1 and α_2 denote the parameters of T_1 and T_2. To simulate X_{jk}, a normal distribution with a mean of 2.0 and a standard deviation of 0.2 is employed (Perera et al., 2016). T_1 and T_2 are created as binary variables with probabilities 0.2 and 0.4, respectively, while the third indicator, T_3, with a probability of 0.6, is considered as the reference level. Using these coefficients and parameters, Y_{1gjk} and Y_{2jk} are simulated by defining the expected values π_{1gjk} and λ_{2jk}, respectively, considering one combination at a time as described below. After generating Poisson data for the second response, the log of the count (LY_{2jk}) is obtained. Hence, in model fitting, the distribution of the survival response is binary, while the distribution of the log count is defined as normal.

The main reason behind using a joint model for the survival and count responses is the dependency between these two responses. To account for the correlation between the responses of the same individual and the correlation within individuals in the same cluster, two levels of random effects are used in the simulation study (Sunethra & Sooriyarachchi, 2020). The random effect for the individuals (second level) is denoted by u_{0jk}, and the random effect for the clusters (third level) is denoted by v_{ok}. One individual represents two responses, which defines the first level of the model. The random effects are generated as $u_{0jk} \sim N(0, \sigma_u^2)$ and $v_{ok} \sim N(0, \sigma_v^2)$. Three different combinations of standard deviations (SD) for u_{ojk} and v_{ok} are considered such that $\sigma_u^2 > \sigma_v^2$: $(\sigma_u, \sigma_v) = (1, 0.75)$, $(1, 0.5)$, and $(0.75, 0.5)$.

The cluster sizes are chosen to be multiples of 5 as it simplifies the implementation of the grouping strategy in simulations. The different combinations of the two main criteria for clusters are as follows: 1) Number of clusters (K) - 3, 5, 7, 10; 2) Number of individuals per cluster (n_k) - 10, 15, 20, 25. If n denotes the total sample size, then $n = n_k \times K$. Thus, the study is conducted considering 4x4 = 16 sample size combinations. It is important to note that there are 16 combinations of sample sizes ($n_k \times K$) for each of the 3 combinations of standard deviations. In total, this results in 3x4x4 = 48 combinations.

To fit the Bayesian marginal models, the following MCMC parameters are used to specify the models: Number of posterior simulation iterations (NMC) = 15000, Thinning rate ($THIN$) = 10, Maximum number of autocorrelation lags ($ACLAG$) = 1000, and Number of tuning iterations (NTU) = 1000. The MCMC sample size (N) retained after thinning is 1500.

3.2 Prior distributions used in the Simulations

To fit hierarchical models using PROC MCMC, prior distributions of the fixed-effect parameters and hyperprior distributions of the random-effect parameters were specified. Normal priors with large constant variances are used for estimating the fixed effects of the joint model as that type of a normal prior is usually considered to be non-informative or weakly-informative. The prior distribution of the variance of random effect is considered as inverse-gamma as using conjugate inverse-gamma distribution for the non-informative prior of variance of normally distributed random effect is more efficient than others. (Chen et al., 2016).

3.3 Proportion of rejections under null hypothesis

The hypotheses of interest are: H_0: The fitted model is adequate vs. H_1: The fitted model is not adequate. As discussed in section 2.5, according to the development of the novel goodness-of-fit (GOF) test, the hypotheses are modified as follows: H_0: $\gamma_{2\_1} = \gamma_{3\_1} = \gamma_{4\_1} = \gamma_{5\_1} = \gamma_{2\_2} = \gamma_{3\_2} = \gamma_{4\_2} = \gamma_{5\_2} = 0$ vs. H_1: At least one γ_{G_i} (for $G = 2, \ldots, 5$ and for $i = 1, 2$) is not equal to zero, where γ_{G_i}'s are the coefficients of the indicator variables in the alternative model (equations (6) and (7)). For each combination, 1000 datasets were generated under the null hypothesis and the rejection proportion of the null hypothesis is calculated as a region of Type I error. To generate the data under this scenario, the priors of γ_{G_i}'s were defined as $\gamma_{G\_i} \sim \text{normal}(0, 10^4)$. The steps were followed as described in section 2.5.

3.3.1 Region of Type I error

Bayesian statistics deals with credible intervals to describe and summarize the uncertainty related to unknown parameters. Within a credible interval, an unobserved parameter value falls with a specific probability, and the generalization to its multivariate problems is known as the credible region. In this study, which involves multiple Bayesian credible intervals, it was observed that the probability of rejecting the null hypothesis (α) is high due to the incorporation of multiple testing. Therefore, the probability outside the credible region is 0.05×4 = 0.2, as each marginal model contains 4 parameters with 95% credibility, according to the *Bonferroni correction* for the issue of multiple comparisons (Bland & Altman, 1995). Thus, to assess the proportion of rejections under H_0, $\alpha = 0.2$ is used. However, due to random variation, it is examined whether the calculated proportion lies within the 95% probability

interval, which is defined as $\alpha \pm Z_{2.5\%}\sqrt{\frac{\alpha(1-\alpha)}{n}}$, where n is the sample size. For $n = 1000$, the 95% probability interval of $\alpha = 0.2$ is $[0.175, 0.225]$.

3.4 Proportion of rejections under alternative hypothesis

It is important to maximize the ability to reject the null hypothesis of a test when the alternative hypothesis is true, which is usually referred to as the power of the test. Therefore, to assess this property of the test, the data are simulated under the alternative hypothesis, and the number of times the test rejects the null hypothesis is calculated. For this step, the priors of γ_{G_i}'s were defined differently from those used in Type I error simulations in section 3.3. Specifically, the prior distribution was defined as $\gamma_{G\_i} \sim \text{normal}(200, 10^4)$, where the mean value is well above zero. This mean value can be any value that is different from zero, and for this study, the power analysis is conducted only for the selected mean value. The other steps were followed as described in section 2.5.

3.5 Simulation results

The simulation results are presented in Table 1 for all 48 combinations of the number of clusters, cluster sizes, and random effect variances. It is important to note that almost all of the 1000 datasets generated for each combination were converged.

The following two points describe the main findings of the simulation study in terms of rejection proportions under H_0.

- When analyzing the results of a small number of clusters (3 and 5), the proportions tend to lie within the limits except only for a few combinations. Moreover, the results are better for the small cluster sizes (10 and 15) even with large numbers of clusters. That is, the combinations of a large number of clusters and large cluster sizes tend to provide bad results.

- Observing the 2nd (1, 0.5) and the 3rd (0.75, 0.5) SD combinations results separately, most of the proportions are smaller than the first SD combination (1, 0.75). As a result, for the number of clusters 7 in the SD combination 2 and 3, more proportions are fallen within the limits.

Table 1: Proportions of rejections under null and alternative hypotheses

No of clusters × Cluster size	**SD combination (σ_u, σ_v)**	**Rejection proportion under H_0**	**Result** (within the limits/ above the upper limit/ below the lower limit)	**Rejection proportion under H_1**
3×10	(1, 0.75)	0.190	Within	0.992
	(1, 0.5)	0.177	Within	0.993
	(0.75, 0.5)	0.175	Within	0.988
3×15	(1, 0.75)	0.205	Within	0.995
	(1, 0.5)	0.180	Within	0.987
	(0.75, 0.5)	0.182	Within	0.985
3×20	(1, 0.75)	0.201	Within	0.988
	(1, 0.5)	0.188	Within	0.993
	(0.75, 0.5)	0.201	Within	0.992
3×25	(1, 0.75)	0.247	Above	0.985
	(1, 0.5)	0.213	Within	0.989
	(0.75, 0.5)	0.212	Within	0.987
5×10	(1, 0.75)	0.197	Within	0.987
	(1, 0.5)	0.188	Within	0.985
	(0.75, 0.5)	0.173	Just Below	0.981
5×15	(1, 0.75)	0.219	Within	0.993
	(1, 0.5)	0.192	Within	0.985
	(0.75, 0.5)	0.192	Within	0.988
5×20	(1, 0.75)	0.223	Within	0.993
	(1, 0.5)	0.217	Within	0.981
	(0.75, 0.5)	0.221	Within	0.979
5×25	(1, 0.75)	0.282	Above	0.985
	(1, 0.5)	0.263	Above	0.979
	(0.75, 0.5)	0.255	Above	0.989
7×10	(1, 0.75)	0.180	Within	0.981
	(1, 0.5)	0.194	Within	0.977
	(0.75, 0.5)	0.167	Below	0.978
7×15	(1, 0.75)	0.234	Above	0.983
	(1, 0.5)	0.224	Within	0.973
	(0.75, 0.5)	0.206	Within	0.965
7×20	(1, 0.75)	0.277	Above	0.967
	(1, 0.5)	0.248	Above	0.977
	(0.75, 0.5)	0.225	Within	0.973
7×25	(1, 0.75)	0.303	Above	0.973
	(1, 0.5)	0.252	Above	0.972
	(0.75, 0.5)	0.280	Above	0.957
10×10	(1, 0.75)	0.191	Within	0.971
	(1, 0.5)	0.192	Within	0.962
	(0.75, 0.5)	0.166	Below	0.952
10×15	(1, 0.75)	0.276	Above	0.967
	(1, 0.5)	0.271	Above	0.958
	(0.75, 0.5)	0.256	Above	0.941
10×20	(1, 0.75)	0.313	Above	0.966
	(1, 0.5)	0.325	Above	0.959
	(0.75, 0.5)	0.309	Above	0.959
10×25	(1, 0.75)	0.344	Above	0.954
	(1, 0.5)	0.307	Above	0.959
	(0.75, 0.5)	0.276	Above	0.946

In the simulations under H_1 (power analysis), almost all the results are above 90%, and it is essential to note that the variations in results are not much influential, as the range of the results is very small (i.e., min = 0.941 and max = 0.995).

- Among all SD combinations, the proportion is largest for the smallest number of clusters (3) and smallest for the largest number of clusters (10). That is, the proportions decrease slightly when the number of clusters increases. However, there is no clear pattern in these proportions with respect to the cluster sizes within each number of clusters.
- For most of the combinations, the results of SD combination 2 and 3 are smaller than the SD combination 1. That is, for low standard deviations of random effects, the power values are low.

4 A real-life example

4.1 The Description of the Dataset

The dataset used for the example has been obtained through a randomized control trial on multi-centred Epilepsy patients (Marson et al., 2002; Sunethra & Sooriyarachchi, 2020). The trial initially has been conducted to compare two antiepileptic treatments at 81 hospitals and follow-up data has been obtained from 1380 epilepsy patients (Marson et al., 2002). The dataset consists of the variables related to patients' demographic information such as age and gender, clinical features such as Epilepsy type and treatment type, statuses of having previous Neurological disorders and previous history of seizures. The selected survival variable is 'time to first seizure after randomization' and the count variable is 'number of seizures experienced by the patient'. The selected variables for the best joint model are defined in section 4.2.

4.1.1 Sample selection

Since the simulations were carried out as to apply the GOF test for small samples having small cluster sizes, a random sample was selected from the original dataset to be matched with the size limits. According to the example a hospital represents a cluster. Among the clusters having more than 10 individuals, 5 clusters/hospitals were selected randomly, as small numbers of clusters provided better results in the simulations. For large clusters, random samples of 25 individuals were obtained. The cluster sizes of the selected random sample of 5 hospitals are 24, 16, 20, 25 and 25. Thus, the total sample size is 108.

4.1.2 Creating a Binary Survival response

The survival time used for the study is 'time to first seizure' experienced by a patient and is a continuous variable as usual. To model this survival time, the

Discrete Time Hazard Model is used where the survival time should be discretized into pre-determined time intervals forming a binary response variable (Goldstein, 2011; Hapugoda & Sooriyarachchi, 2018). Considering literature suggestions, the survival time; 'Time_seizure' was discretized into three time intervals 1-7 days (T_1), 8 days - 1 year (T_2) and > 1 year (T_3). Then an outcome of DTHM was created by linking these time intervals in such a way that the survival time is represented by the binary outcome variable (Hapugoda & Sooriyarachchi, 2018).

4.2 Model fitting and Application of the test

In order to identify the variables that have a significant impact on the responses of interest, separate univariate models for the two responses were fitted including fixed and random effects. Then including the selected significant variables, the joint model is fitted which is used to assess the model adequacy from the novel GOF test. To identify the most significant explanatory variables, the backward elimination method was performed, and the exclusion of variables is decided based on mainly the Deviance Information Criteria (DIC) of the fitted model at each stage. The least significant variables were removed until all the remaining variables seem to be significant according to the DIC and the significance of selected variables is re-checked using the HPD intervals of the model coefficients.

In the joint model, a shared random effect model is used for the marginal log count model (Sunethra & Sooriyarachchi, 2020). For the cluster-specific random effect v_{0k} where $v_{0k} \sim$ normal$(0, \sigma_v^2)$, the shared cluster effects for each marginal model can be written as; for the survival model, $\varphi_{1k} = v_{0k}$ and for the log count model, $\varphi_{2k} = h \cdot v_{0k}$ where h links the two random effects.

The selected joint model is a model including Treatment type = 'treat' (Deferred/Immediate), Status of previous head injury = 'pr_dis_headinj' (Yes/No), Status of previous Meningitis = 'pr_dis_menin' (Yes/No), Status of other Neurological disorder = 'pr_dis_othnd' (Yes/No), Status of previous Febrile convulsions = 'his_sei_febr' (Yes/No), Status of Epilepsy in relatives = 'his_sei_relat' (Yes/No), and Experiencing seizures while asleep = 'asleep' (Yes/No), as exploratory variables.

For the prior distributions, α_l $(l = 0, \ldots, 6)$, β_m $(m = 0, \ldots, 5)$, A_k $(k = 1, \ldots, 4)$, B_k $(k = 1, \ldots, 4) \sim$ normal(0, 10^4), $h \sim$ normal (10,1), $v_{0k} \sim$ normal (0, σ_v^2) where $\sigma_v^2 \sim$ igamma (shape=3, scale=2), the best joint model can be defined as follows.

For Survival $\sim$ Binary(π_{1jk}),

$$\begin{aligned}\text{logit}(\pi_{1jk}) = {} & \alpha_0 + \alpha_1 \text{treat} + \alpha_2 \text{pr\_dis\_othnd} + \alpha_3 \text{his\_sei\_febr} \\ & + \alpha_4 \text{asleep} + \alpha_5 \text{T1} + \alpha_6 \text{T2} \\ & + A_1 \text{Cluster1} + A_2 \text{Cluster2} + A_3 \text{Cluster3} + A_4 \text{Cluster4} + v_{0k}\end{aligned} \tag{8}$$

and for Log(count) $\sim$ Normal(μ_{2jk}, σ^2) where $\sigma^2 = \frac{\exp(\mu_{2jk})}{\text{count}^2}$,

$$\begin{aligned}\mu_{2jk} = \beta_0 &+ \beta_1 \text{pr\_dis\_headinj} + \beta_2 \text{pr\_dis\_menin} + \beta_3 \text{pr\_dis\_othnd} \\ &+ \beta_4 \text{his\_sei\_febr} + \beta_5 \text{his\_sei\_relat} \\ &+ B_1 \text{Cluster1} + B_2 \text{Cluster2} + B_3 \text{Cluster3} + B_4 \text{Cluster4} + h \cdot v_{0k}\end{aligned} \tag{9}$$

For the considered joint Bayesian model above, the fitted values ($\hat{\pi}_{1jk}$ and $\hat{\mu}_{2jk}$) for each observation in the dataset are calculated for survival and log count. Then, by taking one response at a time, the fitted values are sorted and collapsed within each cluster into 5 groups. Four indicators (I_1, I_2, I_3, and I_5) were defined. These five groups are not always allocated as of the same size (roughly equal size) since in the example dataset most of the cluster sizes were not divisible by 5 (Fernando & Sooriyarachchi, 2020). For each response, there will be 4 indicator variables resulting in overall 8 new parameters ($\gamma_{(G\_i)}$ for $G = 2, 3, 4, 5$ and $i = 1, 2$). Then, the model is re-fitted including these indicator variables through the MCMC method.

Table 2: Posterior summaries and intervals of the coefficients of indicators

Parameter (Variable)	**Mean**	**Standard Deviation**	**95% HPD Interval**	
			Lower Bound	**Upper Bound**
$\gamma_{(2_1)}$ ($I_{2\_survival}$)	0.560	0.596	-0.558	1.735
$\gamma_{(3_1)}$ ($I_{3\_survival}$)	-0.262	0.577	-1.321	0.902
$\gamma_{(4_1)}$ ($I_{4\_survival}$)	-0.343	0.413	-1.177	0.475
$\gamma_{(5_1)}$ ($I_{5\_survival}$)	0.008	0.285	-0.502	0.577
$\gamma_{(2_2)}$ ($I_{2\_logcount}$)	0.007	0.010	-0.011	0.025
$\gamma_{(3_2)}$ ($I_{3\_logcount}$)	0.029	0.033	-0.038	0.090
$\gamma_{(4_2)}$ ($I_{4\_logcount}$)	0.004	0.007	-0.012	0.017
$\gamma_{(5_2)}$ ($I_{5\_logcount}$)	0.003	0.010	-0.014	0.024

Finally, by checking the output of the model with indicator variables, the goodness-of-fit of the joint Bayesian model is assessed using the 95% highest posterior density (HPD) intervals of indicator variables. For ease of reference, the output (Table 2) of posterior summaries and HPD intervals is given only for the coefficients ($\gamma_{(G\_i)}$) of the four indicators relevant to the two responses.

Since all the 95% HPD intervals obtained for the coefficients of indicator variables in each marginal model (survival and log count) contain zero, each of the indicator variables is insignificant. This indicates that H_0 should not be rejected. Therefore, based on the novel GOF test, the multilevel joint Bayesian model fitted for the Epilepsy survival and count data is adequate.

5 Discussion and Conclusions

The study mainly focused on assessing the model adequacy of a multilevel joint Bayesian model for survival and count responses. As it was the primary objective of the study, a suitable test was suggested by considering past studies (Hosmer and Lemeshow, 1980; Lipsitz et al., 1996; Perera et al., 2016; and Adhikari & Sooriyarachchi, 2020) and theoretical concepts relevant to Bayesian statistics such as Bayesian credible intervals for assessing the model adequacy.

Secondly, a simulation study was conducted to evaluate the performance of the test by assessing the proportions of rejections under the null hypothesis and the alternative hypothesis of the novel test. The rejection proportions under H_0 were assessed as whether they lie within a probability interval. According to those results, the novel GOF test developed under the Bayesian framework performed well overall for small numbers of clusters, while for large numbers of clusters, the results held within an acceptable region when the cluster size was small. Moreover, there was a decreasing pattern in the probabilities of Type I error when the random-effect standard deviations become small.

Then, a power analysis was conducted under the alternative hypothesis where the datasets were generated using incorrect distributional forms of priors purposely. The proportions under H_1 of small numbers of clusters were slightly higher than the large numbers of clusters for most of the combinations. The proportions tend to decrease when the random-effect standard deviations become small, irrespective of the number of clusters and cluster sizes. Overall, the power of the test is highly satisfactory for all combinations of the number of clusters, cluster sizes, and the random effect variances.

In order to assess how well the test performs in practical scenarios and to have a clear understanding of how the developed test can be applied, the test was illustrated using a real-life Epilepsy dataset. Through the real-world application of the test, it is concluded that the test provides accurate and reliable results for practical applications. Moreover, the test can be used to apply with many explanatory variables and unequal cluster sizes having no difficulties.

Acknowledgement

First and foremost, I would like to extend my extreme gratitude to my supervisor Prof. (Ms.) Roshini Sooriyarachchi, for giving me this invaluable opportunity to work on this study and for the guidance, encouragement and all the support given towards the right path.

I would like to express my appreciation for the generous support given by Dr. A.A. Sunethra in providing a dataset with all the required details to complete the final phase of this study.

References

Adhikari, N., & Sooriyarachchi, M. R. (2020). Developing a goodness of fit test for a joint model of clustered survival and count data. *Communications in Statistics: Simulation and Computation*, 0(0), 1–18. https://doi.org/10.1080/03610918.2020.1825738

Agresti, A. (2003). Categorical data analysis. (Vol. 482), John Wiley & Sons.

Balakrishnan, K., & Sooriyarachchi, M. R. (2018). A goodness of fit test for multilevel survival data. *Communications in Statistics: Simulation and Computation*, 47(1), 30–47. https://doi.org/10.1080/03610918.2016.1186184

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, 18(2), 151–164. https://doi.org/10.1037/a0030642

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. In *Bmj* (Vol. 310, Issue 6973, p. 170). https://doi.org/10.1136/bmj.310.6973.170

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3), 473–514. https://doi.org/10.1214/06-BA117

Chen, F. (2009). Bayesian modeling using the MCMC procedure. *Proceedings of the SAS Global_Forum_2008 Pages1–22*.

Chen, F., Brown, G., & Stokes, M. (2016). Fitting Your Favorite Mixed Models with PROC MCMC_*SAS_Institute*, Pages1–27.

Congdon, P. D. (2010). *Applied Bayesian hierarchical methods*. CRC Press.

Cowling, B. J., Hutton, J. L., & Shaw, J. E. H. (2006). Joint modelling of event counts and survival times. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 55(1), 31–39. https://doi.org/10.1111/j.1467-9876.2005.00529.x

Dunson, D. B., & Herring, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1), 11–25. https://doi.org/10.1093/biostatistics/kxh025

Fernando, G., & Sooriyarachchi, R. (2020). The development of a goodness-of-fit test for high level binary multilevel models. *Communications in Statistics: Simulation and Computation*, 0(0), 1–21. https://doi.org/10.1080/03610918.2019.1700275

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). Bayesian Data Analysis Chapman & Hall. *CRC Texts in Statistical Science*, 105–112.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis*, 1(3), 515–534. https://doi.org/10.1214/06-BA117A

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 339-373.

Goldstein, H. (2011). *Multilevel Statistical Models* (Vol. 922). John Wiley & Sons.

Hapugoda, J. C., & Sooriyarachchi, M. R. (2018a). Joint Modeling of Discrete Time Hazard Model with Poisson Regression Model: A Simulation Study. *Jaffna University International Research Conference, Jaffna, Sri Lanka*, 1–4.

Hapugoda, J. C., & Sooriyarachchi, M. R. (2018b). Joint Modeling of Mixed Responses with Bayesian Modeling and Neural Networks: Performance Comparison with Application to Poultry Data. *Sri Lankan Journal of Applied Statistics*, 19(2), 1. https://doi.org/10.4038/sljastats.v19i2.8019

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

Jayawardena, N. I., & Sooriyarachchi, M. R. (2014). A multilevel bayesian analysis of university entrance eligibility for selected districts in Sri Lanka: Methods and application to educational data. *Journal of the National Science Foundation of Sri Lanka*, 42(1), 23–36. https://doi.org/10.4038/jnsfsr.v42i1.6676

Johnson, V. E. (2004). A Bayesian $\chi 2$ test for goodness-of-fit. *The Annals of Statistics*, 32(6), 2361–2384.

Karunarasan, D., Sooriyarachchi, R., & Pinto, V. (2021). A comparison of Bayesian Markov chain Monte Carlo methods in a multilevel scenario. *Communications in Statistics: Simulation and Computation*, 0(0), 1–17. https://doi.org/10.1080/03610918.2021.1967985

Lemeshow, S., & Hosmer Jr, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, 115(1), 92–106. https://doi.org/10.1093/oxfordjournals.aje.a113458

Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G., Lipsitz, B. S. R., Farber, D., & Fitzmaurice, M. (1996). Response Ordinal for Tests Models Regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 45(2), 175–190.

McNeish, D. M., & Stapleton, L. M. (2016). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychol Review*, 28, 295–314. https://doi.org/10.1007/s10648-014-9287-x

Marson, A. G., Williamson, P. R., Clough, H., Hutton, J. L., & Chadwick, D. W. (2002). Carbamazepine versus Valproate Monotherapy for Epilepsy: A meta-analysis. *Epilepsia*, 43(5), 505–513. https://doi.org/10.1046/j.1528-1157.2002.20801.x

Perera, A. A. P. N. M., Sooriyarachchi, M. R., & Wickramasuriya, S. L. (2016). A Goodness of Fit Test for the Multilevel Logistic Model. *Communications in Statistics: Simulation and Computation*, 45(2), 477–489. https://doi.org/10.1080/03610918.2013.811788

Ranathunga Kapuruge, N. O., & Sooriyarachchi, R. (2017). Multivariate multilevel modeling of age related diseases. *Journal of Modern Applied Statistical Methods*, 16(1), 498–517. https://doi.org/10.22237/jmasm/1493598540

Steenbergen, M. R., & Jones, B. S. (2002). Modelling multilevel data structures. *American Journal of Political Science*, 218-237.

Sunethra, A. A., & Sooriyarachchi, M. R. (2020). A novel method for joint modeling of survival data and count data for both simple randomized and cluster randomized data. *Communications in Statistics - Theory and Methods*, 0(0), 1–23. https://doi.org/10.1080/03610926.2020.1713366