





RESEARCH

Open Access



Zero-augmented beta-prime model for multilevel semi-continuous data: a Bayesian inference

Naser Kamyari¹ , Ali Reza Soltanian^{2*} , Hossein Mahjub³ , Abbas Moghimbeigi⁴  and Maryam Seyedtabib⁵ 

Abstract

Semi-continuous data characterized by an excessive proportion of zeros and right-skewed continuous positive values appear frequently in medical research. One example would be the pharmaceutical expenditure (PE) data for which a substantial proportion of subjects investigated may report zero. Two-part mixed-effects models have been developed to analyse clustered measures of semi-continuous data from multilevel studies. In this study, we propose a new flexible two-part mixed-effects model with skew distributions for nested semi-continuous cost data under the framework of a Bayesian approach. The proposed model specification consists of two mixed-effects models linked by the correlated random effects: Part I) a model on the occurrence of positive values using a generalized logistic mixed model; and Part II) a model on the magnitude of positive values using a linear mixed model where the model errors follow skew distributions including beta-prime (BP). The proposed method is illustrated with pharmaceutical expenditure data from a multilevel observational study and the analytic results are reported by comparing potential models under different skew distributions. Simulation studies are conducted to assess the performance of the proposed model. The DIC₃, LPML, WAIC, and LOO as the Bayesian model selection criteria and measures of divergence used to compare the models.

Keywords: Bayesian framework, Non-negative data, Two-part mixed-effects model, Skew distributions, Pharmaceutical expenditure

Introduction

Semi-continuous data featured with an excessive proportion of zeros and right-skewed positive values arise frequently in health economics and health services research [1]. Examples include alcohol consumption, household-level consumption of food items, medical cost, and substance abuse symptom scales. Statistical models with normality assumptions ignoring the skewness and the spike at zero are not suitable for this

type of data and may lead to substantial bias and incorrect statistical inferences. Two-part (zero-augmented) models, originating in econometrics [2, 3], have been developed extensively in the last three decades to analyse this type of data and have been applied to scientific fields other than economics such as clinical research and health services. In two-part models, we view a semi-continuous variable as the result of two processes: one binomial process determining whether the positive value occurs and one continuous process determining the actual value given it is nonzero. Therefore, a

*Correspondence: soltanian@umsha.ac.ir

² Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Street of Mahdieh, Hamadan, Iran
Full list of author information is available at the end of the article



two-part model consists of two components, with the first component (i.e., Part I) modelling the probability of a response being positive using the probit or logistic regression, and the second component (i.e., Part II) modelling the conditional mean of the positive values (given positive values occurred) using the continuous regression. According to the data structures, various methods have been developed for analysing cross-sectional and longitudinal semi-continuous data [3–7]. Olsen and Schaffer [6] first extended the two-part models developed by Duan et al. [3] and Manning et al. [8] for cross-sectional data to the longitudinal setting by introducing correlated random-effects into the logit and log-normal components, respectively, and applied them to longitudinal alcohol data. In the two-part mixed-effects models, the binomial process is typically modelled with mixed-effects logistic or probit regression, and the continuous process is naturally modelled via linear mixed models (LMMs). The random-effects in the two components are generally assumed to be correlated through a multivariate normal distribution structure. Ignoring the between-component association mistakenly can yield biased estimates in the second part of the model [9]. The correlated random-effects can capture not only the between-component association but also the within-subject correlation among repeated measurements collected from the same individual and nested data. A between-component correlation means that the process giving rise to the positive values is related to the magnitude of the observed value given that a positive response occurred. For example, in a data collection of self-reported daily drinks (DDD) where zero represents no daily drinks and the continuous positive values reflect the mean of drinks per day, a positive correlation suggests that an individual with high odds of drinking tended to drink more alcohol [10].

For the positive part of a semi-continuous variable, LMMs with a normality assumption were used by Husted et al. [11] and Su et al. [9]. However, the positive part of a semi-continuous variable is often right-skewed. The logarithmic transformation was the most commonly used approach to correct the skewness [6, 7] and other monotone increasing functions such as Box-Cox transformation that would make the positive component approximately normal were also explored [12, 13]. The limitations with data transformation in Part II include reduced information, difficulty in interpreting the results and possible heteroscedasticity [13, 14]. An alternative approach is to use generalized linear mixed models (GLMMs) with distributions in the exponential family that can model skewed data, such as Log-Normal, Log-Skew-Normal [15],

Gamma [13], Inverse Gamma, Inverse Gaussian [16], Beta [17], Bridge [18], Generalized Gamma family, and Weibull distributions [19]. It is noted that GLMMs often involve complicated iterative procedures in estimation which may lead to intensive computation burden and non-convergence issues. It would be most effective to use a flexible distribution to model the right skewed positive values in two-part models. Recently, studies have been presented using the beta-prime (BP) distribution to fit long-tail semi-continuous responses [10, 20, 21]. There is very limited research on the application of this skew distribution in a two-part mixed-effects model [22]. This study is an extension of Kamyari et al. (2021), where random effects are added to the linear predictor terms by using real two-level data.

Parameter estimations in two-part modelling could be computationally difficult. For two-part models with independent random-effects, maximum likelihood estimates (MLE) can be derived by fitting separate mixed-effects model to each part [1]. For the correlated two-part mixed-effects model with log-normal distribution on the positive values, Olsen and Shafter [6] and Tooze et al. [7] developed different maximum likelihood approaches. Several authors have proposed Bayesian approaches to fit the two-part models [17, 23–27]. For example, Cooper et al. [23] used a Bayesian approach via Markov Chain Monte Carlo (MCMC) to fit a probit-lognormal correlated two-part model on medical cost data.

As a result, In this study, we propose a two-part mixed-effects model with a logistic mixed model on the occurrence of positive values and a GLMM with BP distribution on the continuous positive values using the Bayesian approach via MCMC procedure with application to a three-level pharmaceutical expenditure data. The data used for this study was extracted from the Iranian pharmaceutical expenditure (PE-2018) survey. The survey was a cross-sectional study that had been conducted by the National Center for Health Insurance Research, Iran Health Insurance Organization. PE-2018 is a dataset of yearly pharmaceutical expenditure per person conducted in 429 cities in Iran.

The rest of the article is organized as follows. At the beginning of **Methods**, we describe the BP regression model. In **Model specification**, we present the two-part mixed effects model for responses with BP distribution. In **Numerical study**, we apply the proposed methodologies to real data and report the analysis results. In **Simulations**, simulation studies are conducted to assess the performance of the proposed models. Finally, we conclude the article with a discussion in **Discussion**.

Methods

Beta-prime regression model

The BP distribution [28, 29] is also known as inverted beta distribution or beta distribution of the second kind, often the model of choice for fitting semi-continuous data where the response variable is measured continuously on the positive real line ($Y > 0$) because of the flexibility it provides in terms of the variety of shapes it can accommodate. The probability density function (PDF) of a BP distributed random variable Y parameterized in terms of its mean μ and a precision parameter ψ is given by

$$f(y|\mu, \psi) = \frac{y^{\mu(\psi+1)-1} (1+y)^{-[\mu(\psi+1)+\psi+2]}}{B(\mu(1+\psi), \psi+2)}, \quad y > 0 \tag{1}$$

where B denote the beta function, $\mu > 0$, $\psi > 0$, $E(Y) = \mu$, and $Var(Y) = (\mu(1+\mu))/\psi$.

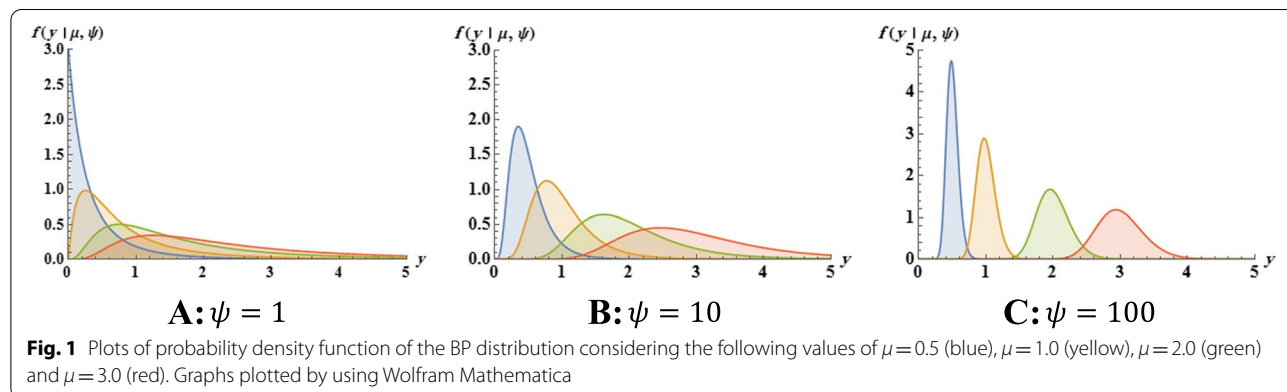
Figure 1 displays some plots of the density function in Eq. (1) for some parameter values. It is evident that the distribution is very flexible and it can be an interesting alternative to other distributions with positive support. Figure 1 shows that for a fixed value of the mean μ , higher values of ψ lead to a reduction of $Var(Y)$, and vice versa. If Y has PDF as in Eq. (1), we denote $Y \sim BP(\mu, \psi)$. Next, to connect the covariate vector X_k , $k = 1, \dots, m$ to the random sample Y_1, Y_2, \dots, Y_m of Y , we use a suitable link function g_1 that maps the mean interval $(0, +\infty)$ onto the real line. This is given as $g_1(\mu_k) = X_k' \beta$, where β is the vector of regression parameters, and the first element of X_k is 1 to accommodate the intercept. The precision parameter ψ_k is either assumed constant [30, 31] or regressed onto the covariates [30, 32] via another link function g_2 , such that $g_2(\psi_k) = Z_k' \gamma$, where Z_k is a covariate vector (not necessarily similar to X_k) and γ is the corresponding vector of regression parameters. Similar to X_k , Z_k also accommodates an intercept. The link functions

$g_1: R^{y>0} \rightarrow R$ and $g_2: R^{y>0} \rightarrow R$ must be strictly monotone, positive and at least twice differentiable, such that $\mu_k = g_1^{-1}(X_k' \beta)$ and $\psi_k = g_2^{-1}(Z_k' \gamma)$, with $g_1^{-1}(\cdot)$ and $g_2^{-1}(\cdot)$ being the inverse functions of $g_1(\cdot)$ and $g_2(\cdot)$, respectively. We can estimate the parameters of the BP regression model defined in Eq. (1) using the `gamlss` function in the R ($\geq 3.3.0$) language [33] with a package of the same name [34].

Model specification

In this section, we present our model for the yearly pharmaceutical expenditure record in three levels. However, our model can be adapted easily to more complicated settings.

In the three-level pharmaceutical expenditure record data, level 3 is the province; level 2 is the city level that nested within province; and level 1 is the subject level that is nested in cities. There are two types of correlations at different levels in the pharmaceutical expenditure data. The first type exists at the province level, where cost records of the same province are correlated. For each subject, there also exists another correlation within each city. Thus, a three-level random effects two-part model is more appropriate for the analysis of the pharmaceutical expenditure data. We are interested in modelling a three-level semi-continuous pharmaceutical expenditure data, characterized by a large portion of zeros and continuous positive values. We define notations as follows. Suppose that we observe cost record y_{ijk} for the k -th subject of city j within province i , where $i = 1, 2, \dots, n$, $j = 1, 2, \dots, n_j$, and $k = 1, 2, \dots, m_{ij}$. The total number of cities is $J = \sum_{i=1}^n n_i$, and the total number of subjects is $N = \sum_{i=1}^n \sum_{j=1}^{n_i} m_{ij}$. Let $\omega_{ijk} = I(y_{ijk} > 0)$ denote the indicator of y_{ijk} being nonzero. Define by X_{ijk} the covariate vectors for the fixed effect. Let p_{1i} and p_{2i} be the correlated random effects in the province level with joint density (p_{1i}, p_{2i}) for parts I and II of our proposed model, respectively.



Similarly, define c_{1ij} and c_{2ij} to be the correlated random effects with joint density (c_{1ij}, c_{2ij}) in the city level. In this paper, we assume that

$$(p_{1i}, p_{2i}) \sim N\left(0, \Sigma_1 = \begin{pmatrix} \sigma_{p_1}^2 & \rho_p \sigma_{p_1} \sigma_{p_2} \\ \rho_p \sigma_{p_1} \sigma_{p_2} & \sigma_{p_2}^2 \end{pmatrix}\right) \text{ and } (c_{1ij}, c_{2ij}) \sim N\left(0, \Sigma_2 = \begin{pmatrix} \sigma_{c_1}^2 & \rho_c \sigma_{c_1} \sigma_{c_2} \\ \rho_c \sigma_{c_1} \sigma_{c_2} & \sigma_{c_2}^2 \end{pmatrix}\right) \tag{2}$$

with Σ_1 and Σ_2 being positive definite matrices. We also assume that (p_{1i}, p_{2i}) and (c_{1ij}, c_{2ij}) are independent for all i 's and j 's. Denote by e_{ijk} the error term for the positive value of Y_{ijk} . We assume that $e_{ijk} \sim N(0, \sigma_e^2)$ is independent of random effects p_{1i}, p_{2i}, c_{1ij} , and c_{2ij} . Define $\tau_{ijk} = P(\omega_{ijk} = 1 | p_{1i}, c_{1ij})$ to be the probability of non-zero value for Y_{ijk} .

To obtain interpretable covariate effects on the marginal mean, we propose the following marginalized two-part model that parameterizes the covariate effects directly in terms of the marginal mean, $\mu_{ijk} = E(Y_{ijk})$, on the original (i.e., untransformed) data scale. The marginalized two-part model with random (cluster) effects for the zero and the continuous components, respectively, specifies the linear predictors.

Part I:

$$\begin{aligned} \logit(\tau_{ijk}) &= \logit(\Pr(Y_{ijk} \neq 0 | p_{1i}, c_{1ij})) = X'_{1ijk} \alpha + p_{1i} + c_{1ij} \\ &= (\alpha_0 + p_{1i} + c_{1ij}) + \sum_{\gamma=1}^r \alpha_\gamma x_{\gamma i} \end{aligned} \tag{3}$$

Part II:

$$L_i = \int \prod_{j=1}^{n_i} \left[\int \exp(l_{ij}^1) \exp(l_{ij}^2) \phi(c_{1ij}, c_{2ij}) dc_{1ij} dc_{2ij} \right] \phi(p_{1i}, p_{2i}) dp_{1i} dp_{2i} \tag{5}$$

$$\begin{aligned} (Y_{ijk} | Y_{ijk} > 0) &\sim \text{Beta Prime}(\mu_{ijk}, \psi) \\ \mu_{ijk} &= E(Y_{ijk} > 0 | p_{2i}, c_{2ij}) = \exp(X'_{2ijk} \beta + p_{2i} + c_{2ij} + e_{ijk}) \\ &= \exp((\beta_0 + p_{2i} + c_{2ij}) + \sum_{\theta=1}^q \beta_\theta x_{\theta i} + e_{ijk}) \end{aligned} \tag{4}$$

where $X_{1N \times (r+1)}$ and $X_{2N \times (q+1)}$ have full rank r and q for the zero and the continuous components, respectively; $\alpha_{(r+1) \times 1}$ and $\beta_{(q+1) \times 1}$ are the corresponding vectors of regression coefficients. As seen in relations (3) and (4), the mixing probability and mean of the component of the continuous parts are linked to the independent variables through logit and logarithmic link functions. The vectors $p_1 = (a_{11}, a_{12}, \dots, a_{1m})'$ and $p_2 = (a_{21}, a_{22}, \dots, a_{2m})'$ denote random effects of the third level in the components of logistic and continuous, respectively, whereas $c_1 = (b_{111}, \dots, b_{11n_1}, \dots, b_{1m1}, \dots, b_{1mnm})'$ and $c_2 = (b_{211}, \dots, b_{21n_1}, \dots, b_{2m1}, \dots, b_{2mnm})'$ are the random effects of the second level. For simplicity of interpretation and mathematical calculations, the random

effects (p_1, p_2) and (c_1, c_2) are assumed to be independent and normally distributed with mean zero and variances $\sigma_{p_1}^2, \sigma_{p_2}^2, \sigma_{c_1}^2$ and $\sigma_{c_2}^2$, respectively [35, 36]. The error terms

$e_{ijk} \sim N(0, \sigma_e^2)$ are also assumed to be normal distribution and independent of the random effects at both levels 2 and 3.

Again and according to the data structure in this study (three-level data), X'_{1ijk} is the vector of covariates for the k -th measurement at the j -th city (level-2) at the i -th province (level-3) for the binary part and X'_{2ijk} is the vector of covariates for the k -th measurement at the j -th city (level-2) at the i -th province (level-3) for the continuous part. The two parts might have common covariates or completely different ones. α is the vector of model coefficients corresponding to the binary part and β is the vector of coefficients corresponding to the continuous part conditional on the values being non-zero. The model can be easily extended to include higher-level random effects.

The conditional PDF for y_{ijk} is expressed as:

$$\begin{aligned} f(y_{ijk} | p_{1i}, p_{2i}, c_{1ij}, c_{2ij}) &= [1 - \tau_{ijk}]^{1-\omega_{ijk}} \times [\tau_{ijk} \times \text{BP}(y_{ijk} | p_{2i}, c_{2ij})]^{\omega_{ijk}} \\ &= \{ [1 - \tau_{ijk}]^{1-\omega_{ijk}} \tau_{ijk}^{\omega_{ijk}} \} \times [\text{BP}(y_{ijk} | p_{2i}, c_{2ij})]^{\omega_{ijk}} \end{aligned}$$

Generally, the estimation of parameters $\alpha, \beta, \psi, \Sigma_1$ and Σ_2 is based on the likelihood function of data given as:

where the log-likelihood for the binary part is

$$\begin{aligned} l_{ij}^1 &= \sum_{k=1}^{m_{ij}} [\omega_{ijk} \log(\tau_{ijk}) + (1 - \omega_{ijk}) \log(1 - \tau_{ijk})] \\ &= \sum_{k=1}^{m_{ij}} [\omega_{ijk} \logit(\tau_{ijk}) + \log(1 - \tau_{ijk})] \end{aligned} \tag{6}$$

and the log-likelihood for the continuous part is

$$l_{ij}^2 = -\frac{1}{2\sigma_e^2} \sum_{k=1}^{m_{ij}} \omega_{ijk} e_{ijk}^2 - \sum_{k=1}^{m_{ij}} \omega_{ijk} \log \sigma_e + \text{constant} \tag{7}$$

With $e_{ijk} = Y_{ijk} - X'_{2ijk} \beta - p_{2i} - c_{2ij}$

In this likelihood function (Eq. 5), $\phi(p_{1i}, p_{2i})$ and $\phi(c_{1ij}, c_{2ij})$ represents the bivariate normal distribution for the random effects with mean vector of zeros and variance-covariance matrix Σ_1 and Σ_2 for zero and non-zero part respectively. As can be seen from Eq. (5), the likelihood function involves the integral with respect to the multivariate normal PDF. Parameter estimation in the

proposed models can be computationally difficult as the likelihood function depends on analytically intractable integrals of a non-linear function with respect to the multivariate normal distribution of random-effects.

Bayesian inferential framework

The parameters in part I and II were individually estimated within a Bayesian inferential framework with MCMC sampling of the posteriors.

Let $\Theta = (\alpha, \beta, \psi, \Sigma_1, \Sigma_2)$ be the collection of unknown population parameters in models (2), (3) and (4). To complete the Bayesian formulation, we specify mutually independent prior distributions for all the unknown parameters as follows:

$$\alpha \sim N_r(\alpha_0, \Lambda_1), \beta \sim N_q(\beta_0, \Lambda_2) \tag{8}$$

$$\psi \sim IG(\alpha, \beta), \Sigma_1^{-1} \sim IW(\Omega_1, \nu_1), \Sigma_2^{-1} \sim IW(\Omega_2, \nu_2)$$

where by considering the information available from literature [1, 16, 37] and range of the parameters Normal (N), Inverse Gamma (IG), and Inverse Wishart (IW) distributions are chosen to simplify computations.

Let the observed data $\mathfrak{D} = \{(\omega_{ij}, y_{ij}, x_{1ij}, x_{2ij}) ; i = 1, \dots, n ; j = 1, \dots, m_i\}$, $f(\cdot)$ be a density function, $f(\cdot | \cdot)$ be a conditional density function and $h(\cdot)$ be a prior density function. We assume that the parameters in Θ are independent of each other; that is:

$$h(\Theta) = h(\alpha)h(\beta)h(\psi)h(\Sigma_1)h(\Sigma_2).$$

After specifying the models for the observed data and prior distributions of the unknown model parameters, we can draw samples for the parameters based on their posterior distributions under the Bayesian framework. Therefore, the joint posterior density of Θ , conditional on \mathfrak{D} , can be determined by

$$f(\Theta | \mathfrak{D}) \propto \left\{ \prod_{i=1}^n \int \prod_{j=1}^{m_i} \left[\int \exp\left(l_{ij}^1\right) \exp\left(l_{ij}^2\right) \phi(c_{1ij}, c_{2ij}) dc_{1ij} dc_{2ij} \right] \phi(p_{1i}, p_{2i}) dp_{1i} dp_{2i} \right\} h(\Theta) \tag{9}$$

The integral in (9) has a high dimension and does not have a closed solution. Analytic approximations to the integrals may not be accurate enough. So, the direct calculation of the posterior distribution of Θ based on the observed data \mathfrak{D} is prohibitive [1]. As an alternative, posterior computation of Θ can proceed using a MCMC procedure via Gibbs sampling or Metropolis-Hastings (M-H) algorithm. While the Gibbs sampler relies on conditional distributions [23, 38–40] the Metropolis-Hastings sampler uses a full joint density distribution to generate a candidate draws [38, 41]. Certainly, there is a large body of work on other computational approaches to sampling (slice sampling, adaptive rejection sampling, Hamiltonian Monte Carlo, etc.); covering such methods is beyond the scope of this study. In an initial review of the software, we

concluded that it would be faster to use the OpenBUGS software. Moreover, due to the high volume of data (calculations) and the time limitation, we could not check the performance of other software. However, OpenBUGS was chosen because of its generality and simplicity. The associated OpenBUGS code is available in Additional file 1: Appendix A.1.

Model complexity and fit

There are a variety of methods to select the model that best fits the data. However, in this research article, we focus on the log pseudo marginal likelihood (LPML) and a modified observed deviance information criterion, denoted here by DIC_3 . In addition, we use of two emerging model selection methods, namely leave-one-out cross-validation (LOO-CV) and widely available information criterion (WAIC), due to their fully Bayesian nature.

The Bayesian Deviance Information Criterion (DIC_3) [42] is used to compare the models fitted. It is defined by

$$DIC_3 = \bar{D}(\theta) + p_D$$

where $\bar{D}(\theta) = -2E\{\log[p(y|\theta)]|y\}$ is the posterior mean deviance taken as Bayesian measure of fit, $p(y|\theta) = \prod_{i=1}^n p(y_i|\theta)$, $E\{\log[p(y|\theta)]|y\}$ is the posterior expectation of $\log[p(y|\theta)]$ and p_D is the effective number of parameters representing model complexity. The DIC_3 is a natural generalization of the Akaike Information Criterion (AIC) [43] and interpreted as a Bayesian measure of fit penalized for increased model complexity. The DIC_3 was developed to solve the problem of determining the ‘effective’ number of parameters (p_D) in complex non-nested hierarchical models and its computation has been

coded into the latest version of WinBUGS (1.4). As for the usual DIC, minimum DIC_3 estimates the model that will make the best short-term predictions [42]. Note that the DIC_3 is only comparable across models with exactly the same observed data.

The LPML [44] is another measure for comparing models that derived from the conditional predictive ordinate (CPO) statistics and is one of the most widely used model selection criteria available in WinBUGS. It is derived from the posterior predictive distribution. For our proposed models, a closed form of the CPO_i is not available. However, a Monte Carlo estimate of CPO_i can be obtained by using a single MCMC sample from the posterior distribution of θ , through a harmonic-mean approximation proposed by [45], as $\hat{CPO}_i = \left\{ \frac{1}{K} \sum_{i=1}^K g(y_i|\theta^{(i)})^{-1} \right\}^{-1}$,

where $\theta^{(1)}, \dots, \theta^{(K)}$ is a post burn-in sample of size K from the posterior distribution from θ , and g is the marginal distribution of Y (integrated over the random effects). A summary statistic of the CPO_i is the LPML, defined by $LPML = \sum_{i=1}^n \log(C\hat{P}O_i)$. Larger values of LPML indicate better fit.

The Watanabe-Akaike (or widely applicable) information criterion (WAIC) [46, 47] is closely related to the more widely known DIC measure, which is based on a deviance. The WAIC is a more fully Bayesian approach for estimating out-of-sample expectation. In general, the WAIC is defined as:

$$WAIC = 2p_{WAIC} - 2LPPD$$

The deviance term in DIC is $\log(p(y|\tilde{\theta}))$ where $\tilde{\theta}$ is a point estimate of θ . For WAIC, this term is replaced by the log pointwise predictive density (LPPD), defined as:

$$LPPD = \sum_{i=1}^n \log \int p(y_i|\theta)p_{post}(\theta)d\theta \approx \sum_{i=1}^n \log \frac{1}{M} \sum_{m=1}^M p(y_i|\theta^{(m)}).$$

Just like DIC, there are variants of WAIC which depend on how p_{WAIC} is defined. Gelman, Hwang, and Vehtari also propose $p_{WAIC2} = \sum_{i=1}^n Var_{post}[\log p(y_i|\theta)]$ as a penalty term, where p_{WAIC2} is “the variance of individual terms in the log predictive density summed over the n data points” [48]. Although DIC is a commonly used measure to compare Bayesian models, WAIC has several advantages over DIC, including that it closely approximates Bayesian cross-validation, it uses the entire posterior distribution and it is invariant to parameterisation [49].

Exact cross-validation requires re-fitting the model with different training sets. Approximate leave-one-out cross-validation (LOO-CV) can be computed easily using importance sampling [50]. The Bayesian LOO estimate of out-of-sample predictive fit is

$$LOO = -2LPPD_{LOO} = -2 \sum_{i=1}^n \log \int p(y_i|\theta)p_{post}(\theta|y_{-i})d\theta$$

where $p_{post}(\theta|y_{-i})$ is the posterior distribution based on the data without the i -th data point. Unlike LPPD that uses data point i for both the computation of posterior distribution and the prediction, here $LPPD_{LOO}$ only uses it for prediction, and hence there is no need for a penalty term to correct the potential bias introduced by using data twice [51].

The question regarding the real data is whether the data had better support a true model. To that end, we fit each model using OpenBUGS and compute DIC and LPML for each. We also export the joint posterior distributions

from OpenBUGS into R and compute WAIC and LOO with “loo” package [52].

Numerical study

Specific models and implementation

In this section, we apply the zero-augmented gamma with random effects and zero-augmented beta-prime with random effects to analyze the multilevel pharmaceutical expenditure dataset previously described, where response (y_{ijk}) is the total pharmaceutical expenditure (\$USD) for all drugs prescribed during a 1 year period related to the subject k ($k=1, \dots, 29,354$) that nested within city j ($j=1, \dots, 429$) that are nested within province i ($i=1, \dots, 31$). From now on, the zero-augmented gamma regression model and zero-augmented beta-prime regression model with multilevel random effects, will be called ZAG-RE model and ZABP-RE model, respectively.

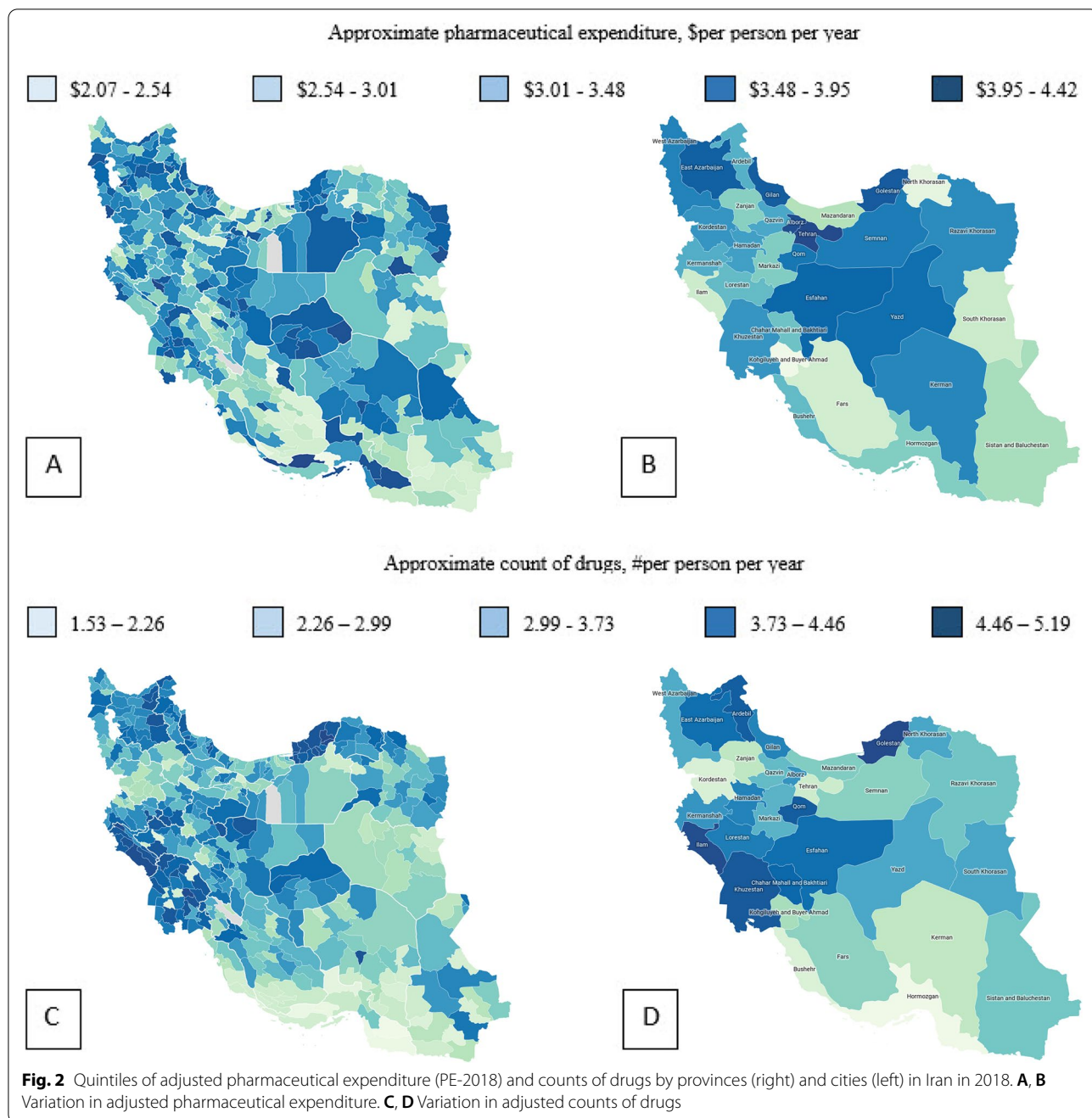
Figure 2(a-d), shows the quintiles of adjusted phar-

maceutical expenditure and counts of drugs by provinces and cities in Iran in 2018. Variation in pharmaceutical expenditure and counts of drugs among clusters (provinces and cities) is well shown. The PE-2018 dataset contains 16.1% of observations with no cost for drugs during the 2018 year. In addition, there is accentuated asymmetry in the empirical distribution of the positive responses, which is confirmed by the sample skewness and sample quartiles (Table 1). These results proposed a skew distribution as a candidate for fitting the pharmaceutical expenditure.

Here we model τ_{ijk} and μ_{ijk} as follows:

$$\begin{aligned} \text{logit}(\tau_{ijk}) &= X'_{1ijk} \times \alpha + p_{1i} + c_{1ij} \text{ and } \log(\mu_{ijk}) \\ &= X'_{2ijk} \times \beta + p_{2i} + c_{2ij} \end{aligned}$$

where X'_{1ijk} and X'_{2ijk} are the matrices of population effects related to subject k , containing an intercept and the following six covariates: Total_k (total inpatients expenditure (\$) per year), ICPD_k (insurance coverage for prescription drugs per year), NDPP_k (number of drugs per prescription), NOP_k (number of prescriptions), Age_k (in year), and Sex_k (1 = male, 0 = female); α and β are coefficient vectors for the mean and zero-part regression, respectively; p_{1i} and p_{2i} be the correlated random effects in the province level with joint density (p_{1i}, p_{2i}) for parts I and II of our model, respectively. Similarly, define c_{1ij} and c_{2ij} to be the correlated random effects with joint density (c_{1ij}, c_{2ij}) in the city level.



In the absence of historical data/experiment, our prior choices follow the specifications described in [Model specification](#). Thus, we consider the following independent (weak) priors for the MCMC sampling:

$$a_r \sim N(0, 10^4), r = 1, 2, 3,$$

$$\beta_q \sim N(0, 10^4), q = 1, \dots, 7,$$

$$\Sigma_1^{-1} \sim IW(0.01I_2, 2), \Sigma_2^{-1} \sim IW(0.01I_2, 2) \text{ and finally}$$

$$\psi \sim IG(0.01, 0.01).$$

Table 1 Percentage of zeros and descriptive statistics of positive expenditure by sex in PE-2018 dataset

	Female (n ₁ = 17,157)	Male (n ₂ = 12,197)	Total (n = 29,354)
Number (%) of zeroes	2576 (15.01%)	2147 (17.60%)	4723 (16.10%)
Mean (USD\$)	2.8306	2.4397	2.6682
Standard deviation	12.4804	10.8963	11.84928
Coefficient of variation	4.4091	4.4662	4.440926
Skewness	19.128	28.746	22.3550
25%	0.3236	0.270	0.3024
50%	0.9108	0.8640	0.8928
75%	2.2392	2.2374	2.2392

PE-2018 data

We generate two parallel independent MCMC runs of size 200,000 – each of them with widely dispersed initial values – and discard the first 100,000 iterations (burn-in samples) for later computing of posterior estimates. We consider a lag of size 100 to eliminate potential problems due to autocorrelation and monitor the convergence of the MCMC chains using trace plots and the R statistic [53], which indicates convergence of about 1. To improve convergence, we divide the response (Y_{ijk}) by 100.

Estimation and model comparison

We consider significant those effects whose 95% equal-tail credible intervals (CI) do not include zero (Table 2 and Fig. 3). Except NDPP and sex in continuous part of ZAG-RE, 95% equal-tail CIs show that, all other variables were significant in two parts of models (Fig. 3). Therefore, the final ZAG-RE and ZABP-RE models have, respectively, the following systematic setting:

$$\begin{aligned}
 \text{logit}(\tau_{ijk}) &= \alpha_0 + \alpha_1 \times NDPP_k + \alpha_2 \times Age_k + \alpha_3 \times Sex_k + p_{1i} + c_{1ij} \text{ and} \\
 \log(\mu_{ijk}) &= \beta_0 + \beta_1 \times Total_k + \beta_2 \times ICPD_k + \beta_3 \times NDPP_k + \beta_4 \times NOP_k + \beta_5 \times Age_k + \beta_6 \times Sex_k + p_{2i} + c_{2ij}.
 \end{aligned}
 \tag{10}$$

The posterior estimates of parameters of model (Eq. 10) shown in Table 2, are quite close in both ZAG-RE and ZABP-RE models while estimates of variance components differ between them. However, 95% equal-tail CI for $\sigma_{p_1p_2}$ and $\sigma_{c_1c_2}$ includes zero in both models, indicating no correlation between variances in level 2 and level 3. Posterior standard deviations of variance components are a bit larger under the ZAG-RE model. Also, it is important note that the meaning of parameter ψ differs between the ZAG-RE and ZABP-RE models. In ZAG-RE model, ψ represents the dispersion parameter, while in

ZABP-RE model, it represents the invariance of Y_{ijk} , conditioned on the random effects.

We use DIC₃, LPML, WAIC, and LOO as the Bayesian model selection criteria and measures of divergence discussed previously to compare the ZAG-RE and ZABP-RE to fit the PE-2018. Except in computational time, The ZABP-RE model performs better according to all other criteria, because it has the smaller DIC₃, WAIC, and LOO-CV and greater LPML (Table 2). Based on those results, we select the ZABP-RE as our best model.

We also conducted a sensitivity analysis on the prior assumptions for the dispersion parameter (ψ) and the fixed effects precision parameter. In particular, we allowed that the dispersion $\psi \sim \text{Gamma}(k, k)$ with $k \in \{0.001, 0.1\}$ and the normal precision on the fixed effects to be 0.1, 0.25 and 0.001. We checked the sensitivity in the posterior estimates of β by changing one parameter at a time and refitting both models.

Although slight changes were observed in parameter estimates and model comparison values, the results appeared to be robust and did not change our conclusions regarding the best model, inference, and sign of the fixed-effects.

From these findings, we further report the results in detail only for the best ZABP-RE model in the following Section.

Results for the ZABP-RE model

We use the ZABP-RE model to interpret parameters effects on the mean of positive expenditure (μ_{ijk}) and the probability of non-cost (τ_{ijk}) by considering individual

Table 2 Bayesian selection criteria and posterior estimates of the ZAG-RE and ZABP-RE models fitted to pharmaceutical expenditure (PE-2018) data

	Criterion	ZAG-RE				ZABP-RE				
	DIC ₃	6754.54				6369.19				
	LPML	- 1124.56				- 1124.28				
	WAIC	6769.13				6398.17				
	LOO-CV	6772.95				6403.25				
	Compute time	4195 s				4338 s				
Model	Parameter	Posterior features								
		Mean	SD	2.5%	97.5%	Mean	SD	2.5%	97.5%	
Zero-part	Intercept	1.645	0.051	1.546	1.744	1.234	0.051	1.522	1.714	
	NDPP	0.117	0.007	0.103	0.132	0.117	0.003	0.093	0.135	
	Age, year	-0.010	0.001	-0.012	-0.009	-0.011	0.001	-0.012	-0.008	
	Male	-0.183	0.032	-0.119	-0.246	-0.184	0.024	-0.116	-0.241	
Continuous-part	Intercept	-2.320	0.050	-2.419	-2.222	-0.988	0.015	-1.069	-0.906	
	Total	0.063	0.002	0.058	0.068	0.013	0.00	0.013	0.013	
	ICPD	-0.061	0.003	-0.066	-0.056	-0.012	0.00	-0.013	-0.12	
	NDPP	-0.001	0.005	-0.010	0.008	-0.038	0.003	-0.045	-0.031	
	NOP	0.110	0.017	0.077	0.142	0.220	0.013	0.194	0.246	
	Age, year	0.003	0.001	0.002	0.005	0.006	0.001	0.004	0.007	
	Male	-0.019	0.025	-0.067	0.029	-0.092	0.023	-0.137	-0.047	
	ψ	0.905	0.014	0.877	0.933	0.499	0.006	0.487	0.512	
	Variance component	σ_{p_1}	0.144	0.015	0.117	0.176	0.148	0.011	0.112	0.153
		σ_{p_2}	0.020	0.011	0.008	0.045	0.019	0.011	0.008	0.041
$\sigma_{p_1 p_2}$		-0.004	0.014	-0.029	0.020	-0.004	0.008	-0.022	0.020	
σ_{c_1}		0.718	0.216	0.404	1.250	0.754	0.214	0.397	1.241	
σ_{c_2}		13.53	16.51	1.097	56.570	10.16	4.891	1.090	35.525	
$\sigma_{c_1 c_2}$		-1.698	1.249	-5.018	0.154	-1.700	0.891	-3.851	0.151	

SD, 2.5 and 97.5% represents standard deviation and percentiles from the posterior distributions of parameters, respectively. Computational time in second
 DIC₃ deviance information criterion, LPML log pseudo marginal likelihood, WAIC watanabe-akaikie information criterion, LOO-CV leave-one-out cross-validation,
 ZAG-RE zero-augmented gamma regression, ZABP-RE zero-augmented beta-prime regression

effects as zero. To measure effects directly on μ_{ijk} and τ_{ijk} , we take the anti-logarithm of $\log(\mu_{ijk})$ and $\text{logit}(\tau_{ijk})$ in Eq. (10), obtaining

Here, α_0 represents the effect of being a female respondent with age and NDPP set to their respective mean. Setting NDPP and age variables to zero – which implies they

$$\mu_{ijk} = \exp(\beta_0 + \beta_1 Total_k + \beta_2 ICPD_k + \beta_3 NDPP_k + \beta_4 NOP_k + \beta_5 Age_k + \beta_6 Sex_k)$$

and

$$\tau_{ijk} = \frac{\exp(\alpha_0 + \alpha_1 \times NDPP_k + \alpha_2 \times Age_k + \alpha_3 \times Sex_k)}{1 + \exp(\alpha_0 + \alpha_1 \times NDPP_k + \alpha_2 \times Age_k + \alpha_3 \times Sex_k)} \tag{11}$$

We use the posterior means in Table 2 as estimates of the parameters. From Eq. (11), parameter β_i represents the rate of change in the logarithm of the mean of the positive expenditure as each of total, ICPD, NDPP, NOP, and age increases one unit. Therefore, increasing NOP variable of subject k in the original scale by one, the $\log(\mu_{ijk})$ increases by 0.22, where $\exp(0.22) = 1.25$ is the increasing value of response variable in its original scale. Parameters $\alpha_0, \alpha_1, \alpha_2$ and α_3 contribute to the calculation of τ_{ijk} in Eq. (11).

are set to their respective mean in the original scale – the probability of no consumption is $1 - \tau_{ijk} = 1 - \exp(1.645) / (1 + \exp(1.645)) = 0.16$ if subject k is female and $1 - \tau_{ijk} = 1 - \exp(1.645 - 0.183) / (1 + \exp(1.645 - 0.183)) = 0.19$ if subject k is male. Overall, females tend to declare a larger expenditure since the estimate of α_3 is negative. Parameters α_1 (0.117) and α_2 (-0.010) represent the effect of NDPP and age variables in $\text{logit}(\tau_{ijk})$. In particular, as NDPP variable increase by one unit, with every additional pharmaceutical

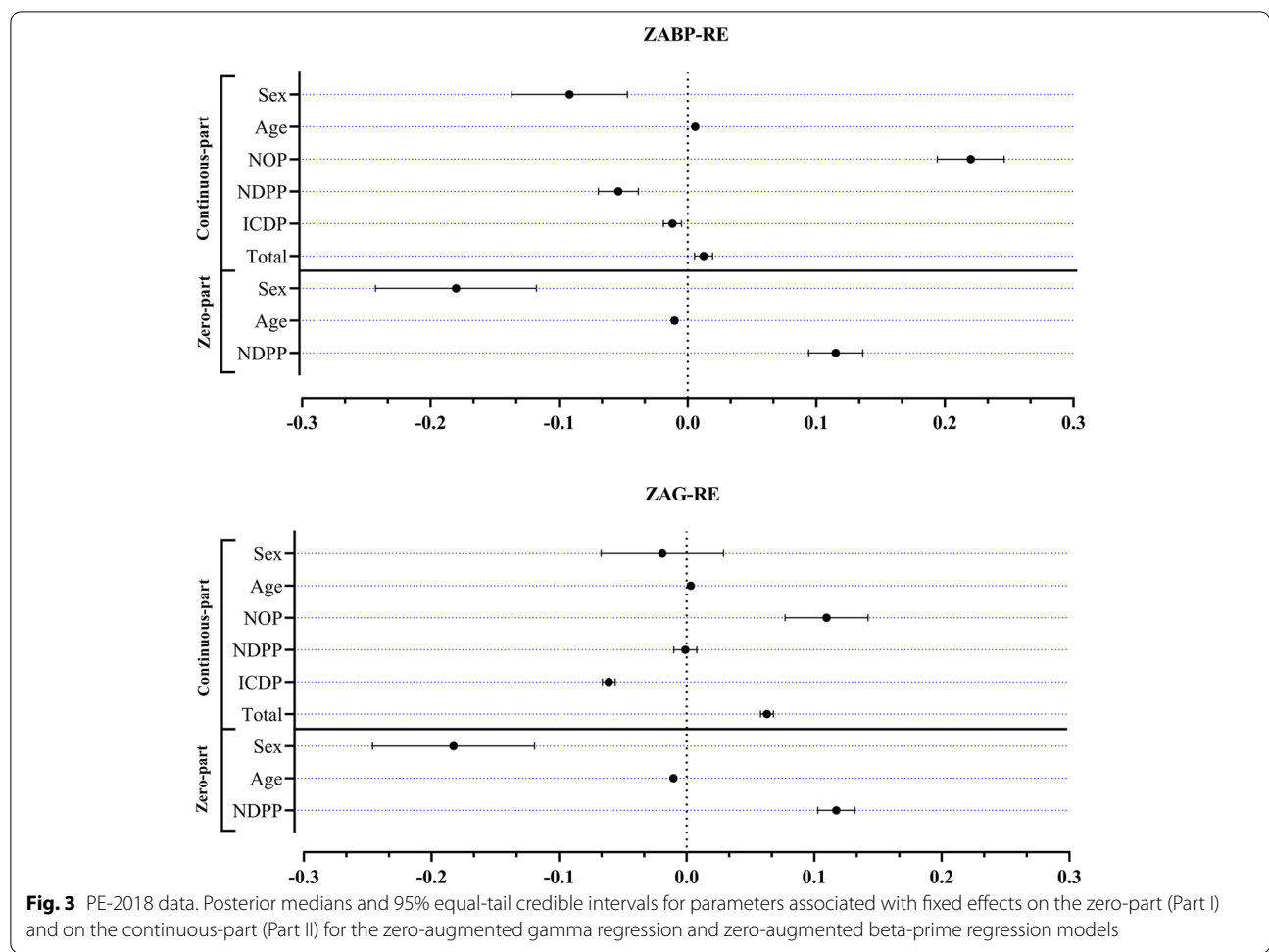


Fig. 3 PE-2018 data. Posterior medians and 95% equal-tail credible intervals for parameters associated with fixed effects on the zero-part (Part I) and on the continuous-part (Part II) for the zero-augmented gamma regression and zero-augmented beta-prime regression models

item, the odds of having a positive expenditure increase by 12.41%. In addition, with each year growing in age, the Odds of having a positive cost decreases by 1%.

In order to evaluate the predictive performance of our best model, we generate 2000 replicates of \mathbf{Y} , say, $\mathbf{Y}^* = (Y^{(1)}, \dots, Y^{(2000)})^T$. The ijk -th element of the l -th replicate $Y_{ijk}^{(l)}$ is generated through the ZABP-RE($\mu_{ijk}^{(l)}, \phi^{(l)}, \tau_{ijk}^{(l)}$) model, where $\mu_{ijk}^{(l)} = \log^{-1} \left(\sum_{\theta=1}^q \beta_{\theta}^{(l)} x_{\theta i} + p_i^{(l)} + c_{ij}^{(l)} \right)$ and $\tau_{ijk}^{(l)} = \text{logit}^{-1} \left(\sum_{\gamma=1}^r \alpha_{\gamma}^{(l)} x_{\gamma i} + p_i^{(l)} + c_{ij}^{(l)} \right)$. The values of $\alpha^{(l)} = (\alpha_0^{(l)}, \dots, \alpha_3^{(l)})$, $\beta^{(l)} = (\beta_0^{(l)}, \dots, \beta_6^{(l)})$ and $\phi^{(l)}$ are post burn-in samples of size 2000 from the posterior distribution of all parameters. Figure 4 (above panel) presents the histogram of PE-2018 placed with the plot of the ZABP-RE and ZAG-RE predictive posterior density. In this figure, it is also quite clear that with a slightly more computational time, the ZABP-RE model provides an adequate fit to the PE-2018 data.

Finally, to evaluate the adequacy of the log-link function used to model the conditional nonzero mean μ , we

follow the suggestion given in [54] as depicted in Fig. 4 (below panel). We divide the values of the linear predictor μ_{ijk} into 10 intervals, with each interval containing a similar number of observations. Then, for each group, we build a boxplot of the posterior predictive mean (black boxplot) and a boxplot of the nonzero observed values (gray boxplot). In Fig. 4, we observe no evidence of link misspecification for the nonzero mean μ_{ijk} , because the shapes of the fitted and observed trends are similar.

Simulations

In this section, we propose a simulation study to illustrate the performance of the proposed method. Our goals for the simulation study were: 1) to investigate the behaviour of Bayesian estimates based on the empirical mean squared error (MSE), relative bias and percentage of times that the 95% credible intervals (CI) contains the true parameter value and 2) to investigate if the DIC₃ and LPML Bayesian criteria properly select the best model. We conduct the simulation study considering 100 datasets generated from ZAG-RE and ZABP-RE models, considering different

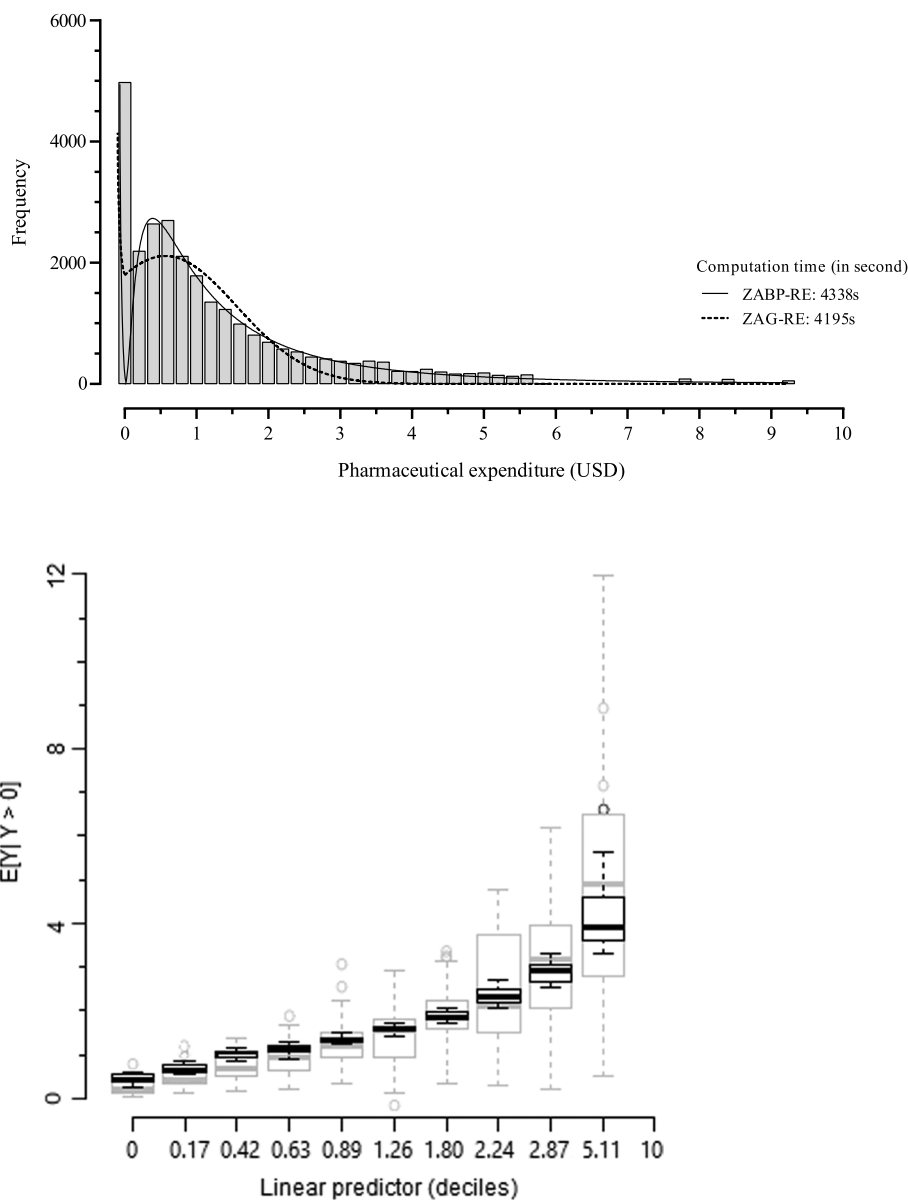


Fig. 4 PE-2018 data. (Above panel) predictive density histogram from pharmaceutical expenditure placed with posterior predictive densities generated using ZABP-RE regression models. (Below panel) adequacy of log link function linear predictor: conditional predictive posterior nonzero mean represented by black boxplots and nonzero observed values by gray boxplot

sample sizes n , say $n=50, 100, 150$ and 200 . For each dataset of size n , we model the location parameter μ_{ijk} through $\log(\mu_{ijk}) = \beta_0 + \beta_1 x_{ijk} + p_i + c_{ij}$ with $p_i \sim N(0, \sigma_p^2)$ and $c_{ij} \sim N(0, \sigma_c^2)$, $i=1, \dots, n, j=1, \dots, n_p, k=1, \dots, m_{ij}$. To keep the simulation simple and fast, τ is considered constant across observations. We generate independent explanatory variables X_{ijk} from a Bernoulli distribution with a parameter equal to 0.8 and set $\beta_0=2, \beta_1=1.5, \tau=0.2, \sigma_p^2=1.8, \sigma_c^2=2.3, \psi=1.0$ for the ZAG-RE model and $\beta_0=2,$

$\beta_1=1.5, \tau=0.2, \sigma_p^2=1.8, \sigma_c^2=2.3, \psi=0.1$ for the ZABP-RE model. We consider the following independent non-informative priors $\beta_k \sim N(0, 100), \psi \sim \text{Gamma}(0.01, 0.01), \sigma_p^2 \sim \text{IGamma}(0.01, 0.01), \sigma_c^2 \sim \text{IGamma}(0.01, 0.01)$ and $\tau \sim U(0, 1)$.

For each dataset of size n , we calculate Bayesian estimates with 500 points from the posterior distribution. These points are based on two parallel independent MCMC runs of size 100,000 each, discarding the first 50,000 points to

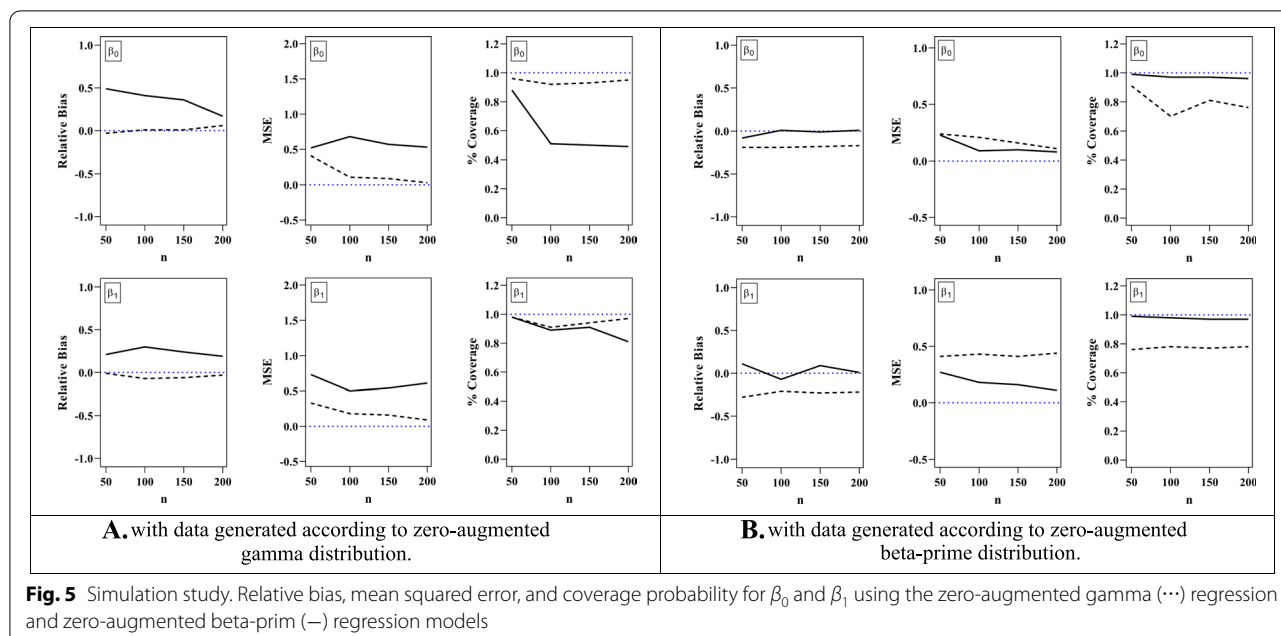


Fig. 5 Simulation study. Relative bias, mean squared error, and coverage probability for β_0 and β_1 using the zero-augmented gamma (···) regression and zero-augmented beta-prim (—) regression models

eliminate the effect of the initial values. To avoid correlation among observations, we consider a thinning of size 100, obtaining 500 points from the posterior distribution.

To study the frequentist properties of Bayesian estimates, we calculate the relative bias, the MSE and the 95% coverage probability (CP). Let $\theta = \{ \alpha, \beta, \psi, \sigma^2 \}$ be the true vector of parameters and θ_s an element of θ . Let $\hat{\theta}_s$ be the posterior mean of 500 points from the posterior distribution of θ_s based on dataset i of size $n, i = 1, \dots, 100, n = 50, 100, 150, 200$. The relative bias, the MSE and the 95% CP for $\hat{\theta}_s$ are defined as follows:

$$\text{Relative bias } (\hat{\theta}_s) = \frac{1}{100} \sum_{i=1}^{100} \left(\frac{\hat{\theta}_{is} - \theta_s}{\theta_s} \right)$$

$$\text{MSE } (\hat{\theta}_s) = \frac{1}{100} \sum_{i=1}^{100} (\hat{\theta}_{is} - \theta_s)^2$$

$$\text{CP}(\hat{\theta}_s) = \frac{1}{100} \sum_{i=1}^{100} I(\theta_s \in [\hat{\theta}_{is,LCL}, \hat{\theta}_{is,UCL}])$$

where I is the indicator function such that θ_s lies in the interval $[\hat{\theta}_{is,LCL}, \hat{\theta}_{is,UCL}]$, with $\hat{\theta}_{is,LCL}$ and $\hat{\theta}_{is,UCL}$ as the estimated lower and upper 95% CIs, respectively. Fig. 5A and B present a visual comparison of the parameters β_0 and β_1 under ZAG-RE and ZABP-RE generated data for varying sample sizes, where the dotted and black lines represent the ZAG-RE and the ZABP-RE fitted model, respectively.

As expected, these figures reveal that if we use a true (ZAG-RE and ZABP-RE) model to fit zero augmented-skew data, relative bias and MSE, for parameters β_0 and β_1 , tend to decrease as sample size increases indicating that the Bayesian estimates possess good consistency properties. In addition, both CP of β_0 and β_1 tend to be around 95% as the sample size increases when the true model is considered. However, when we do not fit the data by their respective true model (model misspecification), the relative bias and the MSE tend to be smaller in ZABP-RE model. Moreover, as expected in both cases we can see that the performance of the CP gets worse when a misspecified model is considered. For the sake of completeness, the MSE, relative bias and CP for all the parameters ($\beta_0, \beta_1, \tau, \sigma^2$) are presented in Table B1 (Additional file 1: Appendix B). It can be seen from this table that the Bayesian estimates of the mixture proportion τ are highly robust to model misspecification, and this behaviour is independent of the sample size. The dispersion parameters (σ^2 and ψ) in both models are not comparable, because they are in different scale. In all results, relative bias of σ_c^2 are greater than σ_p^2 and it shows that the dispersion of responses is more at the level-2 than the level-3. In the comparison of the two models, the differences in MSE values of both σ_p^2 and σ_c^2 are absolutely more in misspecified model. The results of coverage probability of sigmas show that the coverage level in both models are acceptable, however, in true models, the values of the CP are slightly larger than the misspecified model.

Table 3 Summary of model performance in the simulation study

True Model	Criterion	Fitted Model	
		ZAG-RE	ZABP-RE
ZAG-RE	LPML	-4536.87	-6413.12
	DIC ₃	5.12×10^6	5.88×10^7
	WAIC	5.14×10^6	5.89×10^7
	LOO-CV	5.18×10^6	5.93×10^7
	CR%	98%	98%
ZABP-RE	LPML	-9659.82	-8230.09
	DIC ₃	8.45×10^8	5.83×10^8
	WAIC	8.96×10^8	5.89×10^8
	LOO-CV	9.11×10^8	6.13×10^8
	CR%	96%	99%

Simulation study

LPML log pseudo marginal likelihood, DIC₃ deviance information criterion, WAIC watanabe-akaïke information criterion, LOO-CV leave-one-out cross-validation, CR convergence rate, ZAG-RE zero-augmented gamma with random effects, ZABP-RE zero-augmented beta-prime with random effects

Summary of model performance in simulations are presented in Table 3. Table 3 presents the averages of the Bayesian model comparison criteria. We calculate the LPML, DIC₃, WAIC, LOO, and convergence rate using 100 samples of size $n = 100$ each. All criteria favoured the true (simulated) model.

Discussion

In this article, we proposed a Bayesian mixture model with random effects for modelling semi-continuous data augmented by zeros. We suggest the Gamma and BP distributions in continuous part of the models. A simulation study and real data analysis are conducted to compare the ZAG-RE and ZABP-RE on the multi-level semi-continuous data and results demonstrated that the ZABP-RE performs better on the zero-augmented multilevel semi-continuous data.

Our flexible class contains the zero-augmented versions of the two parametric exponential family of distributions, such as Gamma, beta-prime, inverse Gaussian, Weibull, log-normal, and Tweedie. Our model is able to simultaneously accommodate zeros and positive outcomes, right-skewness, within subject correlation because of nested measurements and between-subject heterogeneity. One of the differentials of this study was the inclusion of random effects in the analysis of factors related to semi-continuous data using the beta-prime distribution that were not considered in before studies and statistical packages [10, 20–22], and this is our major contribution.

One of the advantages of the Bayesian approach compared to the classical approach is the estimates in the

part I. Where, the maximum likelihood estimator of a probability of non-zero value, when zero response is observed, does not perform well on the boundary of the parameter space [37]. For a simple BP model, the Maximum Likelihood estimation is available using GAMLSS. However, the MLE results in our data did not reach convergence for some parameters by adding random effects. In this research, using Bayesian statistics with Gibbs and Metropolis-Hasting sampling, this problem is avoided. In the future, it would be interesting to continue the study of various different MCMC methods and hopefully apply them to health cost data.

Simulation studies reveal good consistency properties of the Bayesian estimates as well as high performance of the model selection techniques to pick the appropriately fitted model. We also apply our model to a dataset from yearly pharmaceutical expenditure data conducted in 429 cities in Iran (PE-2018) to illustrate how the procedures can be used to evaluate model assumptions and obtain unbiased parameter estimates. Although our modelling is primarily motivated from the PE-2018, it can be easily applied to other datasets and distributions, because the models considered in this article have been fitted using standard available software packages, like R and OpenBUGS (code available in Additional file 1: Appendix A). This makes our approach easily accessible to practitioners of many fields of research.

Although the zero-augmented positive model considered here has shown great flexibility to deal with zero-augmented clustered data, its robustness can be seriously affected by the presence of heavy tails in the random effects, obscuring important features among individual variation. Liu et al. [13] and Bandyopadhyay et al. [55] proposed a remedy to accommodate skewness in the random effect simultaneously, using skew-normal/independent distributions. We suppose that our method can be used under the zero-augmented positive model and should yield satisfactory results at the expense of additional complexity in implementation. Another useful extension of the proposed model involves the possibility of heteroscedasticity of ψ by allowing the dependence of $g(\psi)$ on covariates, with $g(\cdot)$ being an appropriate link function, as proposed in [56]. An in-depth investigation of such extensions is beyond the scope of the present research article, but it is an interesting topic for further research.

Abbreviations

PE: Pharmaceutical expenditure; BP: Beta-prime; LMMs: Linear mixed models; DDD: Daily drinks data; GLMMs: Generalized linear mixed models; MCMC: Markov Chain Monte Carlo; M-H: Metropolis-Hastings; LPML: Log pseudo marginal likelihood; DIC: Deviance Information Criterion; CPO: Conditional predictive ordinate; WAIC: Watanabe-akaïke information criterion; LOO-CV: Leave-one-out cross-validation; CR: Convergence rate; ZAG-RE: Zero-augmented

gamma regression model with random effects; ZABP-RE: Zero-augmented beta-prime regression model with random effects; NCHIR: National Center for Health Insurance Research.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12874-022-01736-0>.

Additional file 1: Appendix A.1. OpenBUGS code. **Appendix B.** Results of the simulations. **Table B1.** Relative bias, MSE and CP for parameter estimates with different sample sizes for the ZAG-RE and ZABP-RE models.

Acknowledgments

We are thankful to the National Center for Health Insurance Research (NCHIR), which has provided the data for this article. Authors would like to thank the reviewers for the very helpful comments, which lead to considerable improvement of the paper.

Authors' contributions

NK, AS, HM, AM and MS: contributed to the conception, design, and data collection; NK and MS: contributed to the sampling, data gathering, and data assessments. NK, AS, AM, HM and MS: contributed to the statistical analysis and drafting of the manuscript; and AS: supervised the study. All authors read and approved the final version of the manuscript.

Funding

The Abadan University of Medical Sciences, under Grant numbers 1453, supported this work.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author reasonable request.

Declarations

Ethics approval and consent to participate

The Research Ethics Committee of Abadan University of Medical Sciences approved this study with a specific code IRABADANUMS.REC.1401.048. Written informed consent was obtained from all subjects prior to participation in this project.

Consent for publication

Not applicable.

Competing interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Author details

¹Department of Biostatistics and Epidemiology, School of Health, Abadan University of Medical Sciences, Abadan, Iran. ²Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Street of Mahdieh, Hamadan, Iran. ³Research Center for Health Sciences, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran. ⁴Department of Biostatistics and Epidemiology, School of Health, Research Center for Health, Safety and Environment, Alborz University of Medical Sciences, Karaj, Iran. ⁵Department of Biostatistics and Epidemiology, School of Health, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran.

Received: 1 June 2022 Accepted: 27 September 2022

Published online: 02 November 2022

References

- Xing D, Huang Y, Chen H, Zhu Y, Dagne GA, Baldwin J. Bayesian inference for two-part mixed-effects model using skew distributions, with application to longitudinal semicontinuous alcohol data. *Stat Methods Med Res.* 2017;26(4):1838–53.
- Cragg JG. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econom J Econom Soc.* 1971;39(5):829–44.
- Duan N, Manning WG, Morris CN, Newhouse JP. A comparison of alternative models for the demand for medical care. *J Bus Econ Stat.* 1983;1(2):115–26.
- Hall DB, Zhang Z. Marginal models for zero inflated clustered data. *Stat Model.* 2004;4(3):161–80.
- Moulton LH, Curriero FC, Barroso PF. Mixture models for quantitative HIV RNA data. *Stat Methods Med Res.* 2002;11(4):317–25.
- Olsen MK, Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. *J Am Stat Assoc.* 2001;96(454):730–45.
- Tooze JA, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. *Stat Methods Med Res.* 2002;11(4):341–55.
- Manning WG, Morris CN, Newhouse JP, Orr LL, Duan N, Keeler EB, et al. A two-part model of the demand for medical care: preliminary results from the health insurance study. *Health Econ Health Econ.* 1981;137:103–23.
- Su L, Tom BDM, Farewell VT. Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics.* 2009;10(2):374–89.
- Santos-Neto M, Ribeiro-Bezerra T, Bourguignon M, de Castro M. Package "Bpmodel" Title Beta-Prime Regression Model. R package version 1.1.2; 2021.
- Husted JA, Tom BD, Farewell VT, Schentag CT, Gladman DD. A longitudinal study of the effect of disease activity and clinical damage on physical function over the course of psoriatic arthritis: Does the effect change over time? *Arthritis Rheum.* 2007;56(3):840–9.
- Kipnis V, Midthun D, Buckman DW, Dodd KW, Guenther PM, Krebs-Smith SM, et al. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics.* 2009;65(4):1003–10.
- Liu L, Strawderman RL, Cowen ME, Shih Y-CT. A flexible two-part random effects model for correlated medical costs. *J Health Econ.* 2010;29(1):110–23.
- Duan N. Smearing estimate: a nonparametric retransformation method. *J Am Stat Assoc.* 1983;78(383):605–10.
- Smith VA, Preisser JS, Neelon B, Maciejewski ML. A marginalized two-part model for semicontinuous data. *Stat Med.* 2014;33(28):4891–903.
- Rodrigues-Motta M, Galvis Soto DM, Lachos VH, Vilca F, Baltar VT, Junior EV, et al. A mixed-effect model for positive responses augmented by zeros. *Stat Med.* 2015;34(10):1761–78.
- Hatfield LA, Boye ME, Carlin BP. Joint modeling of multiple longitudinal patient-reported outcomes and survival. *J Biopharm Stat.* 2011;21(5):971–91.
- Su L, Tom BDM, Farewell VT. A likelihood-based two-part marginal model for longitudinal semicontinuous data. *Stat Methods Med Res.* 2015;24(2):194–205.
- Jaffa MA, Gebregziabher M, Garrett SM, Luttrell DK, Lipson KE, Luttrell LM, et al. Analysis of longitudinal semicontinuous data using marginalized two-part model. *J Transl Med.* 2018;16(1):1–15.
- Tulupyyev A, Suvorova A, Sousa J, Zelterman D. Beta prime regression with application to risky behavior frequency screening. *Stat Med.* 2013;32(23):4044–56.
- Bourguignon M, Santos-Neto M, de Castro M. A new regression model for positive random variables with skewed and long tail. *Metron.* 2021;79(1):33–55.
- Kamyari N, Soltanian AR, Mahjub H, Moghimbeigi A. Diet, nutrition, obesity, and their implications for COVID-19 mortality: Development of a marginalized two-part model for semicontinuous data. *JMIR Public Health Surveill.* 2021;7(1):e22717.
- Cooper NJ, Lambert PC, Abrams KR, Sutton AJ. Predicting costs over time using Bayesian Markov chain Monte Carlo methods: an application to early inflammatory polyarthritis. *Health Econ.* 2007;16(1):37–56.
- Ghosh P, Albert PS. A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial. *Comput Stat Data Anal.* 2009;53(3):699–706.
- Neelon B, O'Malley AJ, Normand ST. A Bayesian two-part latent class model for longitudinal medical expenditure data: assessing the impact of mental health and substance abuse parity. *Biometrics.* 2011;67(1):280–9.
- Neelon BH, O'Malley AJ, Normand S-LT. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Stat Model.* 2010;10(4):421–39.
- Zhang M, Strawderman RL, Cowen ME, Wells MT. Bayesian inference for a two-part hierarchical model: An application to profiling providers in managed health care. *J Am Stat Assoc.* 2006;101(475):934–45.
- Keeping ES. Introduction to statistical inference. Princeton: D. Van Nostrand Company, Inc.; 1962.

29. McDonald JB. Some generalized functions for the size distribution of income. In: *Modeling income distributions and Lorenz curves*; Springer; 2008. p. 37–55.
30. Bourguignon M, Santos-Neto M, de Castro M. A new regression model for positive data. *arXiv Prepr arXiv180407734*; 2018.
31. Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. *J Appl Stat*. 2004;31(7):799–815.
32. Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol Methods*. 2006;11(1):54.
33. Core Team R. R: a language and environmental for statistical computing. Vienna: R Foundation for Statistical Computing; 2017.
34. Stasinopoulos DM, Rigby RA. Generalized additive models for location scale and shape (GAMLSS) in R. *J Stat Softw*. 2008;23:1–46.
35. Li X, Hedeker D. A three-level mixed-effects location scale model with an application to ecological momentary assessment data. *Stat Med*. 2012;31(26):3192–210.
36. Liu L, Ma JZ, Johnson BA. A multi-level two-part random effects model, with application to an alcohol-dependence study. *Stat Med*. 2008;27(18):3528–39.
37. Rodrigues-Motta M, Forkman J. Bayesian Analysis of Nonnegative Data Using Dependency-Extended Two-Part Models. *J Agric Biol Environ Stat*. 2022;27(2):201–21.
38. Davidian M, Giltinan DM. *Nonlinear models for repeated measurement data*; Routledge; 2017.
39. Huang Y, Wu H. A Bayesian approach for estimating antiviral efficacy in HIV dynamic models. *J Appl Stat*. 2006;33(2):155–74.
40. Sahu SK, Dey DK, Branco MD. A new class of multivariate skew distributions with applications to Bayesian regression models. *Can J Stat*. 2003;31(2):129–50.
41. Ntzoufras I. *Bayesian Modeling Using Winbugs*. Canada: Wiley; 2009.
42. 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):583–639.
43. Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. In: *Selected Papers of Hirotugu Akaike*; 1998. p. 199–213.
44. Carlin BP, Louis TA. *Bayesian methods for data analysis*; CRC Press; 2008.
45. Dey DK, Chen M-H, Chang H. Bayesian Approach for Nonlinear Random Effects Models. *Biometrics*. 1997;53(4):1239 Available from: <http://www.jstor.org/stable/2533493>.
46. Watanabe S, Opper M. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res*. 2010;11(12):1–28.
47. Watanabe S. A widely applicable Bayesian information criterion. *J Mach Learn Res*. 2013;14(27):867–97.
48. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. *Stat Comput*. 2014;24(6):997–1016.
49. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput*. 2017;27(5):1413–32.
50. Gelfand AE. Model determination using sampling-based methods. *Markov Chain Monte Carlo Pract*. 1996;4:145–61.
51. Yong L. LOO and WAIC as model selection methods for polytomous items. *arXiv Prepr arXiv180609996*; 2018.
52. Gabry MJ. Package 'loo'; 2022.
53. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7(4):457–72.
54. Hatfield LA, Boye ME, Hackshaw MD, Carlin BP. Multilevel Bayesian models for survival times and longitudinal patient-reported outcomes with many zeros. *J Am Stat Assoc*. 2012;107(499):875–85.
55. Bandyopadhyay D, Lachos VH, Abanto-Valle CA, Ghosh P. Linear mixed models for skew-normal/independent bivariate responses with an application to periodontal disease. *Stat Med*. 2010;29(25):2643–55.
56. Figueroa-Zúñiga JI, Arellano-Valle RB, Ferrari SLP. Mixed beta regression: A Bayesian perspective. *Comput Stat Data Anal*. 2013;61:137–47.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

