Multilevel Zero-inflated Censored Beta Regression Modeling for Proportions and Rate Data with Extra-zeros

Leili Tapak
Department of Biostatistics, School of Public Health and Modeling of Noncommunicable Disease Research Center, Hamadan University of Medical Sciences, Hamadan, Iran

Omid Hamidi
Department of Science, Hamedan University of Technology, Hamedan, 65155, Iran

Majid Sadeghifar
Department of Science, Bu-Ali Sina University, Hamadan, Iran

Hassan Doosti
Department of Mathematics and Statistics, Sydney, NSW, Australia

Ghobad Moradi
Social Determinants of Health Research Center, Research Institute for Health Development, Kurdistan University of Medical Sciences, Sanandaj, Iran

Corresponding Author: Omid Hamidi
Abstract

**Objectives:** Zero-inflated proportion or rate data nested in clusters due to the sampling structure can be found in many disciplines. Sometimes, the rate response may not be observed for some study units because of some limitations (false negative) like failure in recording data and the zeros are observed instead of the actual value of the rate/proportions (low incidence). In this study, we proposed a multilevel zero-inflated censored Beta regression model that can address zero-inflation rate data with low incidence.

**Methods:** We assumed that the random effects are independent and normally distributed. The performance of the proposed approach was evaluated by application on a three level real data set and a simulation study. We applied the proposed model to analyze brucellosis diagnosis rate data and investigate the effects of climatic and geographical position. For comparison, we also applied the standard zero-inflated censored Beta regression model that does not account for correlation.
Results: Results showed the proposed model performed better than zero-inflated censored Beta based on AIC criterion. Height (p-value <0.0001), temperature (p-value <0.0001) and precipitation (p-value = 0.0006) significantly affected brucellosis rates. While, precipitation in ZICBETA model was not statistically significant (p-value =0.385). Simulation study also showed that the estimations obtained by maximum likelihood approach had reasonable in terms of mean square error.

Conclusions: The results showed that the proposed method can capture the correlations in the real data set and yields accurate parameter estimates.

Keywords: Multilevel modeling; Censored Beta regression; zero-inflation; Mixture distribution; Semicontinuous; Brucellosis.

1. Introduction

Beta regression model is an extension of generalized linear models (GLM) that provides a standard framework for analyzing ratio, rate and proportion data. The main assumption of this model is that there is a continuous, percentage-scaled dependent variable that ranges between 0 to 1 and can be characterized by the Beta distribution [1-4]. Beta distribution characterizes unimodal as well as bimodal densities with varying severity of skewness. This interesting fact provides incredible flexibility for Beta Regression in modeling dependent variables for which normalizing transformations are impossible. In Beta regression, both parameters of the dependent variable including mean and precision (the scaling factor related to variance) are assumed to be associated with explanatory variables [5].

To date, several modeling approaches of the Beta regression have been developed. For example,
a Bayesian approach has been utilized by [6], [7] and [8] to associate the mean and the precision parameters with explanatory variables. A semiparametric Beta regression model has been applied by [9] using penalized splines. Zimprich [10] and Figueroa et al [11] extended Beta regression to longitudinal data analysis through adding a random effect in the mean and precision parameters respectively. Another extension of the Beta regression was used by [12] to analyze time series rate data with correlated errors using Bayesian approach. Other extensions of the Beta regression includes spatial data analysis [13] and spatial analysis of structured additive regression model [14]. In the last recent paper in the Beta regression content [15], the authors focused on random effect models to study spatially correlated rate data to account for the spatial correlation of data in the model. They implemented Bayesian approach for parameter estimation in a two level Beta regression for the correlated data.

There are many situations for which proportion data may contain zeros. In this case, standard form of the Beta regression may fail to model this data. One remedy is to use mixture modeling (a mixture of a Beta distribution on (0, 1) and the Bernoulli distribution which gives non-negative probabilities to 0) to capture the probability mass at 0 [4]. This situation is called zero-inflated Beta distribution [16-18]. “Inflated Beta Regression incorporates the existing Beta distribution with degenerate distributions to model the extreme values, thereby allowing for complete modeling of the entire continuous percentage space” [5].

Often due to the hierarchical design of studies or data collection processes, there are found an intra-class correlation in the data within a cluster. This leads to a correlated data structure (clustered) that is commonly encountered in biomedicine. This characteristic is inherent in health surveys with individuals are nested in clusters, regions or provinces. It is also observed for the observations obtained from the same subject [19]. Ignoring these within-subject (cluster)
dependencies among individuals may lead to misleading statistical inferences because of smaller variance estimates [20, 21]. Multilevel regression modeling is one remedy that can be used in this setting. Several studies extended multilevel zero-inflated models for count data [22, 23]. Schmidt et al utilized a hierarchical mixture Beta dynamic model using a Bayesian approach to model school performance in the Brazilian mathematical Olympiads for public schools [24]. Moreover, a more recent study utilized a multilevel zero-inflated Beta regression model using a Bayesian approach to analyze cluster correlated data with a response variable that take their values in (0,1) interval [25]. The parameter estimations for the later model can be accomplished by separately fitting a binomial regression model for zeros and a Beta model for non-zeros (a two-part model). However, it is also possible to use left-censoring approach and model the zeros as left-censored observations from the Beta distribution. Therefore, it is possible to regard the zeros as having been left censored at the minimum observed rate value (the use of this point maximizes the likelihood) [26]. In such a situation, there can be considered a positive probability for a subject to have a true zero or be a value smaller than the minimum observed value. Example of this is the observed proportion of a disease like brucellosis in different cities of a country where there are zeros that cannot be definitely determined as belonging to true zeros (cities may be unprone to brucellosis due to geographical condition) or low prevalence/incidence (not observed because of low sensitivity of the used tests or inaccessibility to healthcare centers, etc). This type of model is common in statistics. Berk and ALachenbruch [26] applied the same strategy and proposed a zero-inflated left-censored lognormal regression model with random effects.

In this study, a multilevel zero-inflated censored Beta regression model (MLZICBETA) is proposed for proportion data with extra zeros using maximum likelihood estimation approach.
We apply the proposed model to analyze brucellosis diagnosis rate data and investigate the
effects of climatic and geographical position. For comparison, we also apply the standard zero-
inflated censored Beta regression model that does not account for correlation. We show
MLZICBETA is better than zero-inflated censored Beta (ZICBETA). The rest of the paper is
organized as follows. In Section 2, we present a brief review of the standard Beta regression and
its zero-inflated version and then a multilevel zero-inflated censored Beta regression
incorporating random effects to account for data dependency. The details of the proposed
models, the estimation of the parameters, and their variance estimators are also provided. In
Section 3, we report our data analysis for the brucellosis diagnosis rate data. The paper concludes
with a discussion in Section 4.

2. Multilevel ZICBETA regression model

2.1. ZICBETA distribution

Let $w$ denote the proportion/rate/ratio response. The probability density function of a random
variable of $W$ with a Beta distribution can be written as follows:

$$f_w(w; \lambda, \tau) = \frac{\Gamma(\tau)}{\Gamma(\lambda \tau) \Gamma((1 - \lambda) \tau)} w^{\lambda - 1} (1 - w)^{(1 - \lambda) - 1}, 0 < w < 1$$

(1)

where $\tau$ ($\tau > 0$) is the precision parameter and $\lambda$ is the mean of the Beta distribution. The mean
and variance of the Beta random variable ($W$) are given by $E(W | \lambda, \tau) = \lambda$ and

$$Var(W | \lambda, \tau) = \frac{\lambda (1 - \lambda)}{1 + \tau},$$

respectively. According to this relationship, the variance is always
smaller than the mean.
Now, let us consider the response variable $Y$ as a mixture of a degenerated distribution at zero with probability of $p$ and a censored Beta distribution with mean $\lambda$ with probability $1 - p$. The ZICBETA distribution of $y$ can be written:

$$f_r(y; \lambda, \tau) = \begin{cases} 
  p + (1 - p) P(0 < Y < c) & \text{if } y = 0 \\
  (1 - p) f_y(y; \lambda, \tau) & \text{if } c \leq y < 1.
\end{cases}$$

where $I_c(\lambda, \tau)$ is the cumulative probability distribution of Beta distribution and is calculated as follows

$$I_c(\lambda, \tau) = P(Y \leq c) = \frac{\Gamma(\tau)}{\Gamma(\lambda \tau) \Gamma((1 - \lambda) \tau)} \int_0^c y^{\lambda \tau} \Gamma(1 - \lambda) \tau \int_0^y (1 - y)^{(1 - \lambda) \tau - 1} dy, 0 < c < 1$$

Therefore, it can be shown that the mean and variance of $Y$ are given by

$$E(Y) = (1 - p) \lambda$$

and

$$Var(Y) = (1 - p) \frac{(\lambda - c)(1 - \lambda)}{1 + \tau} + p(1 - p) \lambda^2$$

Then the loglikelihood for a sample of $n$ observations censored at $c$ (recorded as zero) and subject to excess zeros is
\[
I(\beta, \gamma; y) = \sum_{i=1}^{n} I(y_i \leq c) \log \left[ p_i(\gamma) + \{1 - p_i(\gamma)\} I(\lambda, \tau) \right] \\
+ \sum_{i=1}^{n} (1 - I(y_i \leq c)) \log \left[ 1 - p_i(\gamma) \right] f_\mu(\beta)
\]

(4)

2.2. Multilevel Zero-Inflated Censored Beta regression

Let \( y_{ijk} \) (for \( i = 1, 2, \ldots, m; j = 1, 2, \ldots, n_i; k = 1, 2, \ldots, n_{ij} \)) be a proportion or rate for the \( k \)th unit in the \( j \)th cluster and \( i \)th province. The total number of clusters is \( n = \sum_{i=1}^{m} n_i \) and the total number of units is \( N = \sum_{i=1}^{m} \sum_{j=1}^{n_i} n_{ij} \). So, the responses of individuals from different provinces (first level) are independent, but they are correlated for those units in the same province. This dependence can be modeled explicitly by considering appropriate random effects in the linear predictor.

ZICBETA model for \( y \in (0,1) \) permits \( \lambda \) to depend on explanatory variables. Then the linear predictors \( \xi_{ijk} \) and \( \eta_{ijk} \) are defined as:

\[
\log \left( \frac{p_{ijk}}{1 - p_{ijk}} \right) = \xi_{ijk} = z_{ijk}^\tau \alpha + w_i + s_{ij}
\]

(5)

\[
\log \left( \frac{\lambda_{ijk}}{1 - \lambda_{ijk}} \right) = \eta_{ijk} = x_{ijk}^\tau \beta + u_i + v_{ij}
\]

(6)

where the covariates \( z_{ijk} \) and \( x_{ijk} \) appearing in the two logistic components are not necessarily the same. The vectors \( w_i \) and \( u_i \) denote the province-specific random effects, whereas \( s_{ij} \) and \( v_{ij} \)
are the random variations at the cluster level. The above equations can be written in vector form as

$$\log \left( \frac{p}{1-p} \right) = \xi = Z\alpha + R_w w + R_s s$$

(7)

$$\log \left( \frac{\lambda}{1-\lambda} \right) = \eta = X\beta + R_u u + R_v v$$

(8)

by considering the random effects in vector form as

$$w = (w_1, \ldots, w_m)^T, \quad u = (u_1, \ldots, u_m)^T,$$

$$s = (s_1, \ldots, s_m, s_{11}, \ldots, s_{2n}, \ldots, s_{m1}, \ldots, s_{mn})^T$$

and

$$v = (v_1, \ldots, v_{1n}, v_{21}, \ldots, v_{2n}, \ldots, v_{m1}, \ldots, v_{mn})^T$$

and design matrices of $Z, X, R_w, R_u, R_v$. For simplicity, we assume that the random effects of $w, s, u$ and $v$ are independent and have normal distribution with mean zero and variances $\sigma_w^2, \sigma_s^2, \sigma_u^2$ and $\sigma_v^2$ respectively.

2.3. Parameter estimation using Expectation Maximization algorithm

The log-likelihood can be maximized by a numerical method like Expectation Maximization (EM) algorithm. We considered the penalized log-likelihood as $l = l_1 + l_2$, where $l_1$ stands for the log-likelihood function when the random effects are assumed conditionally fixed and $l_2$ stands for the $-\log$-likelihood of the random effects. Treating the random effects as parameters leads to the minus $l_2$ to be viewed as a penalty function for the random effects:
Estimating process continues through maximizing $l_1$ by considering variance components as fixed values and updating their values obtained from optimization of $l_2$ using restricted maximum likelihood (REML).

Using the penalized log-likelihood, the complete data log-likelihood ($l_c$) can be constructed as $l_c = l_\xi + l_\eta$ with

$$l_\xi = \sum_{ij_k} \left( u_{ijk} \xi_{ijk} - \log \left( 1 + \exp(\xi_{ijk}) \right) \right)$$

$$- \frac{1}{2} \left[ m \log \left( 2\pi \sigma_u^2 \right) + \sigma_u^{-2} u^T u + n \log \left( 2\pi \sigma_v^2 \right) + \sigma_v^{-2} v^T v \right]$$

$$- \frac{1}{2} \left[ m \log \left( 2\pi \sigma_w^2 \right) + \sigma_w^{-2} w^T w + n \log \left( 2\pi \sigma_s^2 \right) + \sigma_s^{-2} s^T s \right]$$
\[
I_n = \sum_{ijk} \left\{ (1-u_{ijk}) \left[ I(y_{ijk} \leq c) \log(I_c(\lambda, \tau)) + (1-I(y_{ijk} \leq c)) \right] \right. \\
\left. \left\{ \log \left( \frac{\Gamma(\tau)}{\Gamma(\lambda_{ijk} \tau)\Gamma((1-\lambda_{ijk}) \tau)} \right) + (\lambda_{ijk} \tau - 1) \log(y_{ijk}) + ((1-\lambda_{ijk}) \tau - 1) \log(1-y_{ijk}) \right\} \right\} \\
- \frac{1}{2} \left[ m \log(2\pi \sigma_u^2) + \sigma_u^{-2}u^T u + n \log(2\pi \sigma_v^2) + \sigma_v^{-2}v^T v \right]
\]

(12)

where \( u_{ijk} = 1 \) (a latent variable) when \( y_{ijk} \) is zero and \( u_{ijk} = 0 \) if \( y_{ijk} \) is drawn from the censored Beta distribution variable. Therefore, parameter estimation can be performed through EM algorithm. Such decomposition of the complete data log-likelihood enables a convenient method for parameter estimation through expectation maximization method. In the E-step of the EM algorithm, \( u_{ijk} \) is replaced by its conditional expectation \( u_{ijk}^{(g)} \), where \( g \) denotes \( g \)th iteration, under the current values of parameter estimates as follows:

\[
u_{ijk}^{(g)} = \begin{cases} 
1 & \text{if } y = 0 \\
\frac{1}{1 + \exp(-(z_{ijk}^{(g)} \alpha_i^{(g)} + w_i^{(g)} + \delta_j^{(g)})))I_c(\lambda_{ijk}^{(g)}, \tau_{ijk}^{(g)})} & \text{if } c < y < 1 
\end{cases}
\]

(13)

And then in the M-step, (11) and (12) functions are maximized using Newton-Raphson method. M-step of the EM algorithm, estimation of variance components, the scale parameter of NB part, and their corresponding standard errors are given in the Appendix.

3. Application

Brucellosis is a prevalent infectious zoonotic disease that can be transmitted from animals to human directly by contacting with the Brucella carriers or indirectly by contacting with 11
consumption of the unpasteurized dairy products of infected animals (1-3). Symptoms in infected people include fever, sweating and depression as well as arthralgia, and tiredness for weeks or even months (1, 4). Brucellosis can affect public health and economic of a country adversely due to significant human burdens and widespread problems (1, 2). It has been reported that there are more than 500,000 new cases annually worldwide (6, 7). While brucellosis has been eliminated in several industrial countries, it has remained as an important public health menace in Iran with a high annual incidence rate especially in western and northwestern parts of the country (2, 3, 8). Because of the exigent effects of brucellosis, controlling and strategic planning based managing practices are necessary to develop and to improve public health. In the present cross-sectional study, we used a data set on the monthly rate of cases with brucellosis diagnosis (the outcome variable) that were collected for 420 cities (related to 30 provinces; cities are nested in provinces) of Iran from 20 March 2016 to 20 Mar 2017. Data was gathered from the Ministry of Health and Medical Education. We also extracted monthly mean temperature and precipitation values for each city from Iran Meteorological Organization (http://www.irimo.ir/eng/index.php). There were zeros in the rates (37.4%). Here, we encountered with a mixture of zeros and a continuous variable.

Table 1 shows the summary statistics of brucellosis rates by 30 provinces. The majority of the patients were male (58.5%) and aged between 1 and 100 with mean and standard deviation of 38.1 and 17.82 years. About 36.5% of diagnosis had a job connected to livestock and 74.5% of them had a history of using non-pasteurized dairy products. In this cross-sectional study, because of the structure of the data, months were nested within the cities and cities were nested within clusters (provinces). We tried to identify the climatic factors affecting the brucellosis rate using a MLZICBETA regression model. The potential covariates influencing the brucellosis rate were
height of the location, monthly mean temperature (degree of Celsius) and mean monthly precipitation (mm). Table 2 shows the results of MLZICBETA as well as zero-inflated Beta regression models (ZICBETA). We also applied multi-level count regression models including zero-inflated Poisson, zero-inflated negative binomial, zero-inflated generalized Poisson with populations as offset (results were not shown). According to the AIC, our proposed model outperformed other models.

As can be seen, the variance components of random effects were significant in provinces and cities for both parts of the model (zero and Beta parts). Therefore, it can be concluded that the correlations between months in cities and cities that are belong to a province were significant.

After adjusting for the province and cities clustering effects, covariates height (p-value < 0.0001), temperature (p-value < 0.0001) and precipitation (p-value = 0.0006) significantly affect brucellosis cases rates. These results were in concordance with results of other studies [27]. While, precipitation in ZICBETA model was not statistically significant (p-value = 0.385). As the results shown (Table 2) the MLZICBETA had a better AIC compared with the ZICBETA model. According to the results, brucellosis diagnosis rate increases as the temperature increases. This positive relationship was also observed for height. So, cities that are paced in higher locations are associated with greater brucellosis diagnosis rate. In addition, brucellosis diagnosis rate was significantly lower in the area with higher amount of precipitation.

According to the findings, it has been shown that with increasing precipitation amount, the diagnosis rate of brucellosis is reduced. So, in the months and areas with low precipitation the diagnosis rate of brucellosis appears to be more. It seems that in the area with low amount of precipitation the impoverishment of pastures leads to lack of access to nourishing forage with sufficient quality for livestock and because this disease is a viral infection, whenever the
livestock becomes weak due to low protein consumption, the disease incidence increases. In fact, this disease is indirectly related to drought and low precipitation. In recent years, the occurrence of drought phenomenon in semi-arid regions of Iran has caused impoverishment of pasture and deterioration of its quality which has led to protein-energy malnutrition of livestock and consequently the susceptibility of livestock losing in animal husbandries has been increased. Due to the adverse effects of drought and shortage of rainfall on the livestock populations, the favorable conditions of increasing zoonosis diseases in the human population have also been provided. Therefore, there was observed a meaningful relationship between low precipitation amount and the rate of people diagnosed with brucellosis.

Determining the epidemiological status of brucellosis according to special geographical conditions, height and topography in different regions is necessary which requires community-based studies considering the geographical and environmental conditions of each region. According to the results, height has a positive impact of the rate of brucellosis diagnosis. This is because the tribes of nomads and their districts are usually located at altitudes. With the settling of nomads and their livestock at altitudes, the contact between humans and livestock in these areas increases which in turn increases the risk of infection. In Iran, the areas in the highlands have many villages and because of existing abundant pastures animal husbandry has a boom. As a result, high contact with livestock and consumption of dairy products cause an increase in the rate of brucellosis diagnosis in these regions. On the other hand, the lowland regions of Iran are often located in desert and arid areas, because of the lack of sufficient pasture and forages for livestock, there is no livestock (or there is only a few). So, the brucellosis rate in these areas is low.
According to the findings it was indicated that the brucellosis diagnosis rate increases as temperature increases. Increasing temperature in summer provides favorable environmental conditions for the growth of Brucella bacteria and involves human and livestock. On the other hand, as temperature increases, evaporation and transpiration increase and cause the soil to lose its moisture and consequently the quality of the pastures decreases. By diminishing the efficiency in addition to impoverishment of pastures and decreasing the quality of water resources, along with more activity of Brucella bacteria as well as unfavorable environmental conditions at high temperatures for humans and animals lead to an increase in the brucellosis diagnosis rate.

4. Discussion

In this study we proposed a multilevel ZICBETA regression model to analyze hierarchical proportion/rate/ratio data containing extra zeros. This method provides an insight into the source of zeros as well as the apparent heterogeneity. At the same time, it accommodates the within-cluster and within-individual correlations that are introduced by the data structure. Application to the brucellosis diagnosis rate data showed the usefulness of this approach in capturing correlation between study units. In the presence of zeros, the MLZICBETA regression model enables us to draw sensible and valid results and conclusions about significant factors that affect the rate of brucellosis diagnosis and distinguishing those cities with zero rates.

With appropriate modifications, the proposed model can be easily extended to random coefficients setting. It can also be developed for the case of Zero-One-inflated Beta regression model. It is also possible to consider the precision parameter to vary according to the explanatory
variables. This can be done using a logarithm link to associate $\tau$ and potential covariates ($\log(\tau) = \gamma^T X$) with or without random effects. Considering the effects of covariates with an unknown functional structure can be handled using splines which can be another interesting extension of the proposed model. Other extensions include cross-random effects, and non-parametric specification for the distribution of random effects in the two or three parts of the model that are important and interesting areas of future researches.

List of Abbreviations

| Abbreviation | Description |
|--------------|-------------|
| HIV          | human immunodeficiency virus |
| AIDS         | acquired immunodeficiency syndrome |
| MSRIST       | multistate recursively imputed survival trees |
| MSRSF        | multi-state random survival Forest |
| RSF          | random survival Forest |
| RIST         | recursively imputed survival trees |
| Cindex       | Concordance index |
| HAART/ART    | highly active antiretroviral therapy |
| ERMT         | extremely randomized multistate trees |
| OOB          | out-of-bag |
| NRTIs        | nucleoside reverse transcriptase inhibitors |
| TB           | Tuberculosis |

Declarations:

Ethics approval and consent to participate
This study was approved by the Ethics committee of Hamadan University of Medical Science. The data was registry based and innocent. Not applicable.

Consent for publication
Not Applicable.

Availability of data and material
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.
The authors declare no conflict of interest.

This study was partially founded by Hamadan University of medical Science.

LT and OH conceived the research topic. LT, OH, HD and MS explored that idea, performed the statistical analysis and drafted the manuscript. GM provided the data and participated in and drafted the manuscript. All authors read and approved the final manuscript.

We would like to appreciate the Vice-chancellor of Education for technical support and the Vice-chancellor of Research and Technology of Hamadan University of Medical Sciences for their approval and support of this work.

The maximization was performed by the following two sets of Newton–Raphson equations, in view of the orthogonal decomposition of $l_C$

$$
\begin{align*}
\hat{\alpha} & = [\alpha_0] + H_{\alpha,\alpha,\alpha}^{-1} \left[ \frac{\partial \ell_\xi}{\partial \alpha} \right], \\
\hat{w} & = [w_0] + H_{\alpha,\alpha,\alpha}^{-1} \left[ \frac{\partial \ell_\xi}{\partial w} \right], \\
\hat{s} & = [s_0] + H_{\alpha,\alpha,\alpha}^{-1} \left[ \frac{\partial \ell_\xi}{\partial s} \right], \\
\hat{\beta} & = [\beta_0] + H_{\beta,\beta,\beta}^{-1} \left[ \frac{\partial \ell_\eta}{\partial \beta} \right], \\
\hat{u} & = [u_0] + H_{\beta,\beta,\beta}^{-1} \left[ \frac{\partial \ell_\eta}{\partial u} \right], \\
\hat{v} & = [v_0] + H_{\beta,\beta,\beta}^{-1} \left[ \frac{\partial \ell_\eta}{\partial v} \right],
\end{align*}
$$

where $\{\alpha_0, w_0, s_0\}$ and $\{\beta_0, u_0, v_0\}$ stand for the initial values of the parameters and in the iterative method of maximization they are replaced by their updated estimates in each step of iteration. The first and second derivatives of the likelihood of $\ell_\xi$ for the logistic part of the model are given as follows

$$
\frac{\partial \ell_\xi}{\partial \alpha} = Z^T \frac{\partial \ell_\xi}{\partial \xi}, \quad \frac{\partial \ell_\xi}{\partial w} = R_w \frac{\partial \ell_\xi}{\partial \xi} - \sigma_w^{-2} w, \quad \frac{\partial \ell_\xi}{\partial s} = R_s \frac{\partial \ell_\xi}{\partial \xi} - \sigma_s^{-2} s,
$$

$$
H_{\alpha,\alpha,\alpha} = \begin{bmatrix} Z^T & -\frac{\partial^2 \ell_\xi}{\partial \xi \partial \xi^T} & Z \\ R_w^T & & R_w \\ R_s^T & & R_s \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma_w^{-2} I_m & 0 \\ 0 & 0 & \sigma_s^{-2} I_s \end{bmatrix}
$$

where
The first and second derivatives of the likelihood of $\ell_{\eta}$ for the Beta part of the model are given as follows:

$$\frac{\partial \ell_{\eta}}{\partial \alpha} = X^T \frac{\partial \ell_{\eta}}{\partial \eta}, \quad \frac{\partial \ell_{\eta}}{\partial u} = R_u^T \frac{\partial \ell_{\eta}}{\partial \eta} - \sigma_{\eta}^2 \eta, \quad \frac{\partial \ell_{\eta}}{\partial v} = R_v^T \frac{\partial \ell_{\eta}}{\partial \eta} - \sigma_v^2 v,$$

$$H_{\mu, u, v} = \begin{bmatrix} X^T u \\ R_u^T \\ R_v^T \end{bmatrix} \left( -\frac{\partial^2 \ell_{\eta}}{\partial \eta \partial \eta^T} \right) \begin{bmatrix} A & R_u & R_v \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \sigma_u^2 I_m & 0 \\ 0 & 0 & \sigma_v^2 I_s \end{bmatrix},$$

where

$$\frac{\partial \ell_{\eta}}{\partial \eta_{ijk}} = \left( 1 - u_{ijk} \right) \left( 1 \left( y_{ijk} \leq c \right) - \left( 1 - 1 \left( y_{ijk} \leq c \right) \right) \right) \left( \psi \left( \frac{\tau \exp(\eta_{ijk})}{1 + \exp(\eta_{ijk})} \right) - \psi \left( \frac{\tau}{1 + \exp(\eta_{ijk})} \right) + \frac{\tau \exp(\eta_{ijk})}{1 + \exp(\eta_{ijk})} \log \left( \frac{w_{ijk}}{1 - w_{ijk}} \right) \right),$$

$$\frac{\partial^2 \ell_{\eta}}{\partial \eta \partial \eta^T} = \text{Diag} \left[ \left( 1 - u_{ijk} \right) \left( 1 \left( y_{ijk} \leq c \right) - \left( 1 - 1 \left( y_{ijk} \leq c \right) \right) \right) \left( -\psi' \left( \frac{\tau \exp(\eta_{ijk})}{1 + \exp(\eta_{ijk})} \right) - \psi' \left( \frac{\tau}{1 + \exp(\eta_{ijk})} \right) + \tau \log \left( \frac{w_{ijk}}{1 - w_{ijk}} \right) \frac{\exp(\eta_{ijk})}{1 + \exp(\eta_{ijk})^2} \right) \right].$$

Where $\psi(x) = \frac{d}{dx} \log(\Gamma(x))$ is digamma function and $\psi'(x) = \frac{d^2}{dx^2} \log(\Gamma(x)) = \int_0^\infty t^{-3} e^{-t} (\ln(t))^2 dt$.

**Variance components and standard errors**

1. Estimation of the variance of the random effects requires the calculation of the information matrix. The expectations of the second derivatives are:
Then, the information matrix is given by

\[
\frac{\partial^2 \ell}{\partial \xi \partial \bar{\xi}^T} = \text{Diag} \left( I \left( y = 0 \right) \right) \left( \frac{\partial I_c(\lambda, \tau)}{\partial \eta} \frac{\exp(\xi)}{I_c(\lambda, \tau) + \exp(\xi)} \right) - \frac{\exp(\xi)}{(1 + \exp(\xi))^2},
\]

\[
\frac{\partial^2 \ell}{\partial \xi \partial \eta^T} = \text{Diag} \left( I \left( y = 0 \right) \right) \left( \frac{\partial I_c(\lambda, \tau)}{\partial \eta} \frac{\exp(\xi)}{I_c(\lambda, \tau) + \exp(\xi)} \right)
\]

\[
\frac{\partial^2 \ell}{\partial \eta \partial \eta^T} = \text{Diag} \left( I \left( c < y < 1 \right) \right) \left( \frac{\partial^2 I_c(\lambda, \tau)}{\partial \eta \partial \eta^T} \left( I_c(\lambda, \tau) + \exp(\xi) \right) - \left( \frac{\partial I_c(\lambda, \tau)}{\partial \eta} \right)^2 \right) - \frac{\exp(\xi)}{(1 + \exp(\xi))^2},
\]

\[
\frac{\partial \ell_1}{\partial \eta} = \frac{\partial I_c(\lambda, \tau)}{\partial \eta} \frac{\exp(\xi)}{I_c(\lambda, \tau) + \exp(\xi)}
\]

\[
\frac{\partial \ell_1}{\partial \xi} = \frac{\partial I_c(\lambda, \tau)}{\partial \xi} \frac{\exp(\xi)}{I_c(\lambda, \tau) + \exp(\xi)}
\]

\[
\frac{\partial \ell_1}{\partial \bar{\xi}} = \frac{\partial I_c(\lambda, \tau)}{\partial \bar{\xi}} \frac{\exp(\xi)}{I_c(\lambda, \tau) + \exp(\xi)}
\]

\[
(1-1(y_{ijk} < c)) \left( -\psi \left( \frac{\tau \exp(\eta_{ijk})}{1 + \exp(\eta_{ijk})} \right) - \psi \left( \frac{\tau}{1 + \exp(\eta_{ijk})} \right) + \frac{\tau \exp(\eta_{ijk})}{(1 + \exp(\eta_{ijk}))^2} \log \left( \frac{w_{ijk}}{1 - w_{ijk}} \right) \right),
\]
Each element of $H=(H_{ij})$ is a block matrix corresponding to random effects. Then the REML estimates of the variance components are estimated as follows:

$$
\hat{\sigma}_w^2 = \left[ \hat{w}^T \hat{w} + tr(H_{22}) \right] / m,
\hat{\sigma}_s^2 = \left[ \hat{u}^T \hat{u} + tr(H_{33}) \right] / m,
\hat{\sigma}_y^2 = \left[ \hat{v}^T \hat{v} + tr(H_{66}) \right] / n.
$$

Finally, the square root of diagonal elements of the block matrices $H_{11}$ and $H_{44}$ provides the respective standard errors for the estimates of regression coefficients $\alpha$ and $\beta$ and asymptotic variances of the estimators in the variance component is obtained from the inverse of the REML information matrix [28, 29] as follows:

$$
V = \text{var}\begin{bmatrix} \hat{\sigma}_w^2 \\ \hat{\sigma}_s^2 \\ \hat{\sigma}_u^2 \\ \hat{\sigma}_v^2 \end{bmatrix} = 2\left( a_{ij} \right)
$$

Where its elements are as follows:
\[ a_{11} = \sigma_w^{-4} \text{tr} \left( I_m - \frac{h_{23}}{\sigma_w^2} \right)^2, \quad a_{12} = a_{21} = \sigma_w^{-4} \sigma_s^{-4} \text{tr} \left( H_{23} H_{32} \right), \]

\[ a_{13} = a_{31} = \sigma_w^{-4} \sigma_u^{-4} \text{tr} \left( H_{25} H_{52} \right), \quad a_{14} = a_{41} = \sigma_w^{-4} \sigma_v^{-4} \text{tr} \left( H_{26} H_{62} \right), \]

\[ a_{22} = \sigma_s^{-4} \text{tr} \left( I_n - \frac{h_{33}}{\sigma_s^2} \right)^2, \quad a_{23} = a_{32} = \sigma_s^{-4} \sigma_u^{-4} \text{tr} \left( H_{35} H_{53} \right), \]

\[ a_{24} = a_{42} = \sigma_s^{-4} \sigma_v^{-4} \text{tr} \left( H_{36} H_{63} \right), \quad a_{33} = \sigma_u^{-4} \text{tr} \left( I_m - \frac{h_{53}}{\sigma_u^2} \right)^2, \]

\[ a_{34} = a_{43} = \sigma_u^{-4} \sigma_v^{-4} \text{tr} \left( H_{56} H_{65} \right), \quad a_{44} = \sigma_v^{-4} \text{tr} \left( I_m - \frac{h_{66}}{\sigma_v^2} \right)^2, \]

The standard errors of variances of \( \sigma_w^2, \sigma_s^2, \sigma_u^2 \) and \( \sigma_v^2 \) are the square roots of the diagonal elements of matrix \( V^{-1} \).
References

1. Paolino P: Maximum likelihood estimation of models with beta-distributed dependent variables. Political Analysis 2001, 9(4):325-346.

2. Kieschnick R, McCullough BD: Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Statistical modelling 2003, 3(3):193-213.

3. Ferrari S, Cribari-Neto F: Beta regression for modelling rates and proportions. Journal of Applied Statistics 2004, 31(7):799-815.

4. Ospina R, Ferrari SL: Inflated beta distributions. Statistical Papers 2010, 51(1):111.

5. Swearingen CJ, Castro MM, Bursac Z: Inflated beta regression: Zero, one and everything in between. In: SAS Institute, SAS Global Forum: 2012; 2012.

6. Cepeda E, Gamerman D: Bayesian methodology for modeling parameters in the two parameter exponential family. Revista Estadística 2005, 57(168-169):93-105.

7. Simas AB, Barreto-Souza W, Rocha AV: Improved estimators for a general class of beta regression models. Computational Statistics & Data Analysis 2010, 54(2):348-366.

8. Smithson M, Verkuilen J: A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological methods 2006, 11(1):54.

9. Branscum AJ, Johnson WO, Thurmond MC: BAYESIAN BETA REGRESSION: APPLICATIONS TO HOUSEHOLD EXPENDITURE DATA AND GENETIC DISTANCE BETWEEN FOOT-AND-MOUTH DISEASE VIRUSES. Australian & New Zealand Journal of Statistics 2007, 49(3):287-301.

10. Zimprich D: Modeling change in skewed variables using mixed beta regression models. Research in Human Development 2010, 7(1):9-26.

11. Figueroa-Zúñiga JI, Arellano-Valle RB, Ferrari SL: Mixed beta regression: A Bayesian perspective. Computational Statistics & Data Analysis 2013, 61:137-147.

12. Ferreira G, Figueroa-Zúñiga JI, de Castro M: Partially linear beta regression model with autoregressive errors. Test 2015, 24(4):752-775.

13. Cepeda-Cuervo E, Núñez-Antón V: Spatial double generalized beta regression models: extensions and application to study quality of education in Colombia. Journal of Educational and Behavioral Statistics 2013, 38(6):604-628.

14. Gholizadeh K, Mohammadzadeh M, Ghayyomi Z: Spatial Analysis of Structured Additive Regression and Modeling of Crime Data in Tehran City Using Integrated Nested Laplace Approximation. Journal of Statistical Sciences 2013, 7(1):103-124.

15. Kalhori L, Mohammadzadeh M: Spatial Beta Regression Model with Random Effect. Journal of Statistical Research of Iran JSRI 2017, 13(2):215-230.

16. Ospina R, Ferrari SL: A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis 2012, 56(6):1609-1623.

17. Peng X, Li G, Liu Z: Zero-inflated beta regression for differential abundance analysis with metagenomics data. Journal of Computational Biology 2016, 23(2):102-110.

18. Liu F, Eugenio EC: A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. Statistical methods in medical research 2016:0962280216650699.
19. Kong M, Xu S, Levy SM, Datta S: GEE type inference for clustered zero-inflated negative binomial regression with application to dental caries. *Computational statistics & data analysis* 2015, **85**:54-66.

20. Hedeker D, Gibbons RD: *Longitudinal data analysis*, vol. 451: John Wiley & Sons; 2006.

21. Fitzmaurice GM, Laird NM, Ware JH: *Applied longitudinal analysis*, vol. 998: John Wiley & Sons; 2012.

22. Moghimbeigi A, Eshraghian MR, Mohammad K, Mcardle B: Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics* 2008, **35**(10):1193-1202.

23. Lee AH, Wang K, Scott JA, Yau KK, McLachlan GJ: Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical methods in medical research* 2006, **15**(1):47-61.

24. Schmidt AM, de Moraes CP, Migon HS: A Hierarchical Mixture Beta Dynamic Model of School Performance in the Brazilian Mathematical Olympiads for Public Schools. *Chilean Journal of Statistics (ChJS)* 2017, 8(1).

25. Schmidt AM, de Moraes CP, Migon HS: A Hierarchical Mixture Beta Dynamic Model of School Performance in the Brazilian Mathematical Olympiads for Public Schools. 2017, 8(1).

26. Berk KN: *Repeated measures with zeros*. 2002, **11**(4):303-316.

27. Ahmadkhani M, Alesheikh AA: Space-time analysis of human brucellosis considering environmental factors in Iran. *Asian Pacific Journal of Tropical Disease* 2017, **7**(5):257-265.

28. McGilchrist C, Yau K: The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models. *Communications in statistics-theory and methods* 1995, **24**(12):2963-2980.

29. Wang K, Yau KK, Lee AH, McLachlan GJ: Multilevel survival modelling of recurrent urinary tract infections. *Computer methods and programs in biomedicine* 2007, **87**(3):225-229.
Table 1. Descriptive statistics of Brucellosis rate in different provinces.

| Province clusters                  | Number of Cities | Mean       | Standard deviation | Minimum   | Maximum   |
|------------------------------------|------------------|------------|--------------------|-----------|-----------|
| East Azerbaijan                    | 20               | 0.000214   | 0.000499           | 0.000000  | 0.003154  |
| West Azerbaijan                    | 17               | 0.000090   | 0.000125           | 0.000000  | 0.000919  |
| Ardabil                            | 10               | 0.000075   | 0.000112           | 0.000000  | 0.000733  |
| Isfahan                            | 23               | 0.000051   | 0.000093           | 0.000000  | 0.000632  |
| Alborz                             | 6                | 0.000017   | 0.000047           | 0.000000  | 0.000282  |
| Ilam                               | 10               | 0.000055   | 0.000144           | 0.000000  | 0.001113  |
| Bushehr                            | 8                | 0.000007   | 0.000017           | 0.000000  | 0.000105  |
| Tehran                             | 16               | 0.000016   | 0.000045           | 0.000000  | 0.000431  |
| Chaharmahal and Bakhtiari           | 9                | 0.000103   | 0.000352           | 0.000000  | 0.002101  |
| South Khorasan                     | 11               | 0.000099   | 0.000198           | 0.000000  | 0.001235  |
| Razavi Khorasan                    | 28               | 0.000193   | 0.000324           | 0.000000  | 0.002251  |
| North Khorasan                     | 7                | 0.000096   | 0.000136           | 0.000000  | 0.000629  |
| Khuzestan Province                 | 27               | 0.000024   | 0.000080           | 0.000000  | 0.000949  |
| Zanjan                             | 8                | 0.000231   | 0.000291           | 0.000000  | 0.001201  |
| Semnan                             | 7                | 0.000026   | 0.000046           | 0.000000  | 0.000268  |
| Sistan and Baluchestan             | 18               | 0.000033   | 0.000135           | 0.000000  | 0.001384  |
| Fars                               | 35               | 0.000068   | 0.000112           | 0.000000  | 0.001113  |
| Ghazvin                            | 6                | 0.000172   | 0.000397           | 0.000000  | 0.002439  |
| Kurdistan                          | 10               | 0.000154   | 0.000204           | 0.000000  | 0.001326  |
| Kerman                             | 23               | 0.000048   | 0.000097           | 0.000000  | 0.000728  |
| Kermanshah                         | 14               | 0.000107   | 0.000133           | 0.000000  | 0.000762  |
| Kohgiluyeh and Boyer-Ahmad         | 8                | 0.000033   | 0.000065           | 0.000000  | 0.000465  |
| Golestan                           | 14               | 0.000076   | 0.000147           | 0.000000  | 0.001384  |
| Gilan                              | 13               | 0.000025   | 0.000061           | 0.000000  | 0.000485  |
| Lorestan                           | 12               | 0.000398   | 0.000973           | 0.000000  | 0.007105  |
| Mazandaran                         | 20               | 0.000051   | 0.000110           | 0.000000  | 0.000818  |
| Markazi                            | 12               | 0.000103   | 0.000183           | 0.000000  | 0.001001  |
| Hormozgan                          | 8                | 0.000066   | 0.000191           | 0.000000  | 0.000118  |
| Hamadan                            | 9                | 0.000163   | 0.000195           | 0.000000  | 0.001158  |
| Yazd                               | 10               | 0.000024   | 0.000046           | 0.000000  | 0.000217  |
Table 2. Parameter estimates and standard errors for multilevel ZICBETA and ZIBETA regression models

|                        | Logistic part |                  |                  | Beta part        |                  |                  |
|------------------------|---------------|------------------|------------------|------------------|------------------|------------------|
|                        | Estimate      | Standard Error   | P-value          | Estimate         | Standard Error   | P-value          |
| Intercept              | -10.391       | 3.717            | 0.005            | 4.780            | 0.080            | <0.0001          |
| Precipitation          | -0.024        | 0.048            | 0.620            | -0.051           | 0.015            | 0.0006           |
| Temperature            | -0.342        | 0.047            | <0.0001          | 0.188            | 0.014            | <0.0001          |
| Height                 | -0.537        | 0.158            | <0.0001          | 0.0007           | 0.369            | 0.065            |
| Sigma (Province)       | 0.266         | 0.040            | <0.0001          | 0.334            | 0.067            | <0.0001          |
| Sigma (Cities)         | 1.461         | 0.088            | <0.0001          | 0.804            | 0.033            | <0.0001          |
| \( \tau \)             | 168.680       | 5.045            | <0.0001          |                  |                  |                  |
| -2log-likelihood       | -18262        |                  |                  |                  |                  |                  |
| AIC                    | -18236        |                  |                  |                  |                  |                  |
| BIC                    | -18218        |                  |                  |                  |                  |                  |

|                        | ZICBETA       |                  |                  |                  |                  |                  |
|                        | Estimate      | Standard Error   | P-value          | Estimate         | Standard Error   | P-value          |
| Intercept              | -0.543        | 0.030            | <0.0001          | 4.214            | 0.023            | <0.0001          |
| Precipitation          | -0.004        | 0.033            | 0.890            | 0.015            | 0.017            | 0.385            |
| Temperature            | 0.02167       | 0.034            | 0.522            | 0.054            | 0.018            | 0.002            |
| Height                 | -0.464        | 0.032            | <0.0001          | 0.149            | 0.018            | <0.0001          |
| \( \tau \)             | 40.938        | 1.255            | <0.0001          |                  |                  |                  |
| -2log-likelihood       | -14475        |                  |                  |                  |                  |                  |
| AIC                    | -14457        |                  |                  |                  |                  |                  |
| BIC                    | -14399        |                  |                  |                  |                  |                  |
