Augmented mixed beta regression models for periodontal proportion data

Diana M. Galvis[a] Dipankar Badyophadyay[b,*] Victor H. Lachos[a]

[a]*Departamento de Estatística, Universidade Estadual de Campinas, Brazil*

[b]*Division of Biostatistics, University of Minnesota, Minneapolis, MN*

Abstract

Continuous (clustered) proportion data often arise in various domains of medicine and public health where the response variable of interest is a proportion (or percentage) quantifying disease status for the cluster units, ranging between zero and one. However, due to the presence of relatively disease-free as well as highly diseased subjects in any study, the proportion values can lie in the interval $[0, 1]$. While Beta regression can be adapted for assessing covariate effects here, it's versatility is often challenged due to the presence/excess of zeros and ones because the Beta support lies in the interval $(0, 1)$. To circumvent this, we augment the probabilities of zero and one with the Beta density, controlling for the clustering effect. Our approach is Bayesian with the ability to borrow information across various stages of the complex model hierarchy, and produces a computationally convenient framework amenable to available freeware. The marginal likelihood is tractable, and can be used to develop Bayesian case-deletion influence diagnostics based on q-divergence measures. Both simulation studies and application to a real dataset from a clinical periodontology study quantify the gain in model fit and parameter estimation over other adhoc alternatives, and provide quantitative insight into assessing the true covariate effects on the proportion responses.

Keywords Augmented Beta; Bayesian; Beta density; Kullback-Leibler divergence; Periodontal disease.

1 Introduction

Clinical studies often generate proportion data where the response of interest is continuous and confined in the interval $(0, 1)$, such as percentages, proportions, fractions and rates Kieschnick & McCullough (2003). Examples include proportion of nucleotides that differ for a given sequence or gene in foot-and-mouth disease Branscum *et al.* (2007), the percent decrease in glomerular filtration rate at various follow-up times since baseline Song & Tan (2000), etc. With the usual fidelity towards the usual Gaussian assumptions for model errors, one might here be tempted enough to fit a linear regression model to assess the response-covariate relationship Qiu *et al.* (2008). However, this leads to misleading conclusions by ignoring the range constraints in the responses. The logistic-normal model of Aitchison (1986) which assumes normal distribution for

*Division of Biostatistics, University of Minnesota SPH, A452 Mayo MMC 303, 420 Delaware Street SE, Minneapolis, MN 55455, USA E-mail: dbandyop@umn.edu

logit-transformed proportion responses might provide a computationally convenient framework here, however it suffers from an interpretation problem given that the expected value of response is not a simple logit function of the covariates. In this context, the Beta Regression (BR), proposed by Ferrari & Cribari-Neto (2004) can accomplish a direct modeling scheme of covariates under a generalized linear model (GLM) specification, leading to easy interpretation. The Beta density Johnson *et al.* (1994) is extremely flexible, can take on a variety of shapes, and accounts for the heteroscedasticity, non-normality, and skewness in the data. The BR model consider a specific re-parametrization of the associated Beta density parameters, and connects the covariates with the mean and precision of the density through appropriate link functions. Despite its versatility, its potential is limited for proportion responses with support in $(0, 1)$.

The motivating data example for this paper comes from a clinical study conducted at the Medical University of South Carolina (MUSC) to determine the periodontal health status of Type-2 diabetic Gullah-speaking African Americans Fernandes *et al.* (2006). For assessing dental health, the clinical attachment level (or, CAL), a clinical marker of periodontal disease (PD), is measured at each of the 6 sites of a tooth clustered within a subject. Considering this clustering, the underlying statistical question is to estimate the functions that model the dependence of the 'proportion of diseased sites corresponding to a specific tooth-type (represented by incisors, canines, pre-molars and molars)' with the covariables. Figure 1 (left panel) plots the raw (unadjusted) density histogram of the proportion responses in our data, packed over subjects and tooth-types. The responses lie in the closed interval $[0, 1]$ where 0 and 1 represent 'completely disease free', and 'highly diseased' cases, respectively. Although BR might be applicable here post (adhoc) re-scaling Smithson & Verkuilen (2006) of the data from $[0, 1]$ to the interval $(0, 1)$, various limitations are observed working on a transformed scale, viz., (i) loss of information on an underlying data generation scheme, (ii) component-wise transformation might not yield a convenient joint modeling framework, (iii) parameters might loose interpretability on a transformed scale, and (iv) transformations may not be universal Lachos *et al.* (2011). Adhoc re-scalings might provide a nice working solution for small proportions of 0's and 1's, sensitivity towards parameter estimation can be considerable with higher proportions. Hence, from a practical perspective, there is a necessity to seek an appropriate theoretical model that avoids data transformations, yet capable of handling the challenges the data presents. This inefficiency is only escalated due to the presence of additional clustering in the data, as in our case. To circumvent this, we propose an efficient generalized linear mixed model (GLMM) framework by augmenting the probabilities of occurrence of zeros and ones to the BR model via a zero-and-one-augmented Beta (ZOAB) random effects (ZOAB-RE) model, that can accommodate the subject-level clustering.

Staring with Ferrari & Cribari-Neto (2004), there has been various specifications of the BR model. The BR model of Ferrari & Cribari-Neto (2004) re-parameterizes the Beta density parameters and connects the data covariates to the response mean via a logit link, assuming that the data precision is constant (nuisance) across all observations. This was subsequently modified linking the covariates to the dispersion parameter via the variable dispersion BR model by Smithson & Verkuilen (2006). Very recently, Verkuilen & Smithson (2012) used Gauss-Hermite quadrature for calculating ML estimates and a Gibbs sampler for Bayesian estimation in the context of BR models for correlated proportion data. Also, Figueroa-ZúñIga *et al.* (2012) presents a Bayesian approach for the correlated BR model through Gibbs samplers, and uses the deviance information criterion (DIC) Carlin & Louis (2008), Expected-AIC (EAIC) and Expected-BIC (EBIC) for model selection. However, to the best of our knowledge, there are no studies that utilizes a Bayesian paradigm to model clustered (correlated) proportion data where the proportions lie in the interval [0,1]. Our proposition 'augments' point masses at zero and one to a continuous (Beta) density that does not

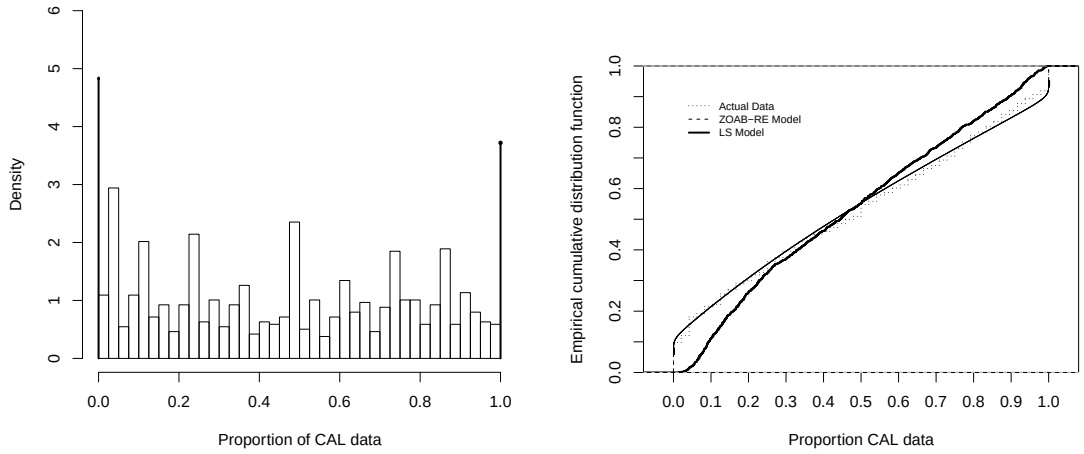

Figure 1: Periodontal proportion data. Left panel plots the (raw) density histogram packed over subjects and tooth-types. Right panel presents the empirical cumulative distribution function of the real data, and that obtained after fitting the ZOAB-RE model and the LS model.

include zero and one in its support in a similar spirit of Hatfield *et al.* (2012). In addition, following the pioneering work of Cook (1986), we develop case-deletion and local influence diagnostics to assess the effect of outliers on the parameter estimates. Our approach is Bayesian, with the ability to borrow information across various stages of the complex model hierarchy, and produces a computationally convenient framework amenable to available freeware like `WinBUGS`.

The rest of the article proceeds as follows. After a brief introduction to the BR model, Section 2 introduces the ZOAB-RE model, and develops the Bayesian estimation scheme. Section 3 presents Bayesian model selection tools and related case influence diagnostics. Section 4 applies the proposed ZOAB-RE model to the motivating data, and uses Bayesian model selection and outlier detection tools to select the best model. It also summarizes and discusses the estimation of the fixed effects, and other related model parameters. Section 5 presents a simulation study assessing the finite sample performance of our method with another competing transformation-based model. Conclusions and future developments appear in Section 6.

2 Statistical Model and Bayesian Inference

2.1 Beta Regression model

The Beta distribution is often the model of choice for fitting continuous data restricted in the interval (0,1) due to the flexibility it provides in terms of the variety of shapes it can accommodate. The probability density function of a Beta distributed random variable Y parameterized in terms of its mean $\mu(0<\mu<1)$ and a precision parameter $\phi(\phi>0)$ is given by

$$f(Y=y|\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\ 0<y<1, \tag{1}$$

where $\Gamma(.)$ denotes the gamma function, $E(Y)=\mu$, and $\text{Var}(Y)=\frac{\mu(1-\mu)}{1+\phi}$. Therefore, for a fixed

value of the mean μ, higher values of ϕ leads to a reduction of Var(Y) and conversely. If Y has pdf as in (1), we write $Y \sim \text{beta}(\mu\phi;(1-\mu)\phi)$. The BR model can be defined as follows. Let $Y_1,\ldots,Y_n$ be n independent random variables with the covariate vector $\mathbf{x}_i$ such that $Y_i \sim \text{beta}(\mu_i\phi_i;(1-\mu_i)\phi_i)$. Next, to connect $\mathbf{x}_i$ to the response Y, we use a suitable link function g_1 that maps the mean interval (0,1) onto the real line. This is given as $g_1(\mu_i) = \mathbf{x}_i^\top\beta$, where β is the vector of regression parameters, and the first element of $\mathbf{x}_i$ is 1 to accommodate the intercept. The precision parameter ϕ_i is either assumed constantFerrari & Cribari-Neto (2004), or regressed upon the covariatesSmithson & Verkuilen (2006) via another set of link functions g_2, such that $g_2(\phi_i) = \mathbf{z}_i^T\alpha$, where $\mathbf{z}_i$ is a covariate vector (not necessarily similar to $\mathbf{x}_i$) and α is the corresponding vector of regression parameters. Similar to $\mathbf{x}_i$, $\mathbf{z}_i$ also accommodates an intercept. Both g_1 and g_2 are strictly monotonic, and twice differentiable. Choices of g_1 includes the logit specification $g_1(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$, the probit function $g_1(\mu_i) = \Phi^{-1}(\mu_i)$ where $\Phi(.)$ is the standard normal density, the complimentary log-log function $g_1(\mu_i) = \log\{-\log(1-\mu_i)\}$ among others, and for g_2, the log function $g_2(\phi_i) = \log(\phi)$, the square-root function $g_2(\phi_i) = \sqrt{\phi_i}$, and the identity function $g_2(\phi_i) = \phi_i$ (with special attention to the positivity of the estimates) Simas *et al.* (2010). Estimation follows via either the (classical) maximum likelihood (ML) route Ferrari & Cribari-Neto (2004); Smithson & Verkuilen (2006) through Gauss-Hermite quadratures available in the *betareg* library in `R` Zeileis *et al.* (2010), or Bayesian Branscum *et al.* (2007) through Gibbs sampling.

2.2 Zero-and-one augmented Beta random effects model

The BR model described above only applies for observations that are independent, and moreover it is suitable only for responses lying in $(0,1)$. However, for our motivating periodontal proportion dataset, the responses pertaining to a particular subject are clustered in nature, and lie bounded in $[0,1]$. We now develop a ZOAB model to address both the bounded support problem, and the clustered data. Our proposition comprises a three-part mixture distribution, with degenerate point masses at 0 and 1, and a Beta density to have the support of $Y_i \in 0 \cup 1 \cup (0,1)$. Thus, $Y \sim \text{ZOAB}(p_0,p_1,\mu_i,\phi)$, if the density of Y_i, $i=1,\ldots,n$, follows:

$$f(Y_i = y_i|p_0,p_1,\mu_i,\phi) = \begin{cases} p_0 & \text{if } y_i = 0 \\ p_1 & \text{if } y_i = 1 \\ (1-p_0-p_1)f(Y_i = y_i|\mu_i,\phi) & \text{if } y_i \in (0,1), \end{cases} \tag{2}$$

where $p_0 \geq 0$ denotes the probability $Y_i = 0$, $p_1 \geq 0$ denotes the probability $Y_i = 1$, $0 \leq p_0 + p_1 \leq 1$ and $f(y|\mu,\phi)$ is given in (1). The mean and variance of Y_i is given by

$$\begin{aligned} E[Y_i] &= \gamma_i = (1-p_0-p_1)\mu_i + p_1, \\ \text{Var}(Y_i) &= p_1(1-p_1) + (1-p_0-p_1)\left[\frac{\mu_i(1-\mu_i)}{1+\phi} + (p_0+p_1)\mu_i^2 - 2\mu_i p_1\right]. \end{aligned}$$

The ZOAB-RE model is now defined as follows. Let $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ be n independent continuous random vectors, where $\mathbf{Y}_i = (y_{i1},\ldots,y_{in_i})$ is the vector of length n_i for the sample unit i, with the components $y_{ij} \in [0,1]$. Next, the covariates can be regressed over a suitably transformed μ_{ij} as above, such that

$$G(E[Y_{ij}|\mathbf{b}_i]) = g_1(\mu_{ij}) = \mathbf{X}_i^\top\beta + \mathbf{Z}_i^\top\mathbf{b}_i, \tag{3}$$

where $\mathbf{X}_i$ is the design matrix of dimension $p \times n_i$ corresponding to the vector of fixed effects $\beta = (\beta_1,\ldots,\beta_p)^\top$, and $\mathbf{Z}_i$ is the design matrix of dimension $q \times n_i$ corresponding to REs vector $\mathbf{b}_i = (b_{i1},\ldots,b_{iq})^\top$. Choice of link functions for g_1 remain the same as in Subsection 2.1, however,

we prefer to use the logit link for our model development. In this study, we are more interested in the 'at-risk' population, i.e. where the proportion responses are in $(0,1)$. Our motivating dataset on Gullah-speaking African-Americans is homogenous in nature (everyone has Type-2 diabetes), hence the extreme (boundary) responses (0's and 1's) representing 'no-disease' and 'completely-diseased' states are assumed to be representative of two distinct, but homogenous categories. For efficient parameterization, we choose to use p_0, p_1, and ϕ (the dispersion parameter) as constants in further model development, however relaxing those to be to be p_{0ij}, p_{1ij}, and ϕ_{ij}, and estimation using priors on these parameters, or regressing those to model covariates through appropriate link functions (say, logit/probit/cloglog for p_{0ij} and p_{1ij}, and log for ϕ_{ij}) is certainly possible. Thus, we define our ZOAB-RE model as $Y_{ij} \sim \text{ZOAB-RE}(p_0, p_1, \mu_{ij}, \phi)$ $i = 1,\ldots,n,\ j = 1,\ldots,n_i$.

2.3 Data likelihood

Let $\Omega = (p_0, p_1, \phi, \beta)$ denote the parameter vector in this ZOAB-RE model. The primary goal here is to estimate Ω, and to derive inference on β adjusting for the effects of clustering. Considering the RE design matrix $\mathbf{Z}_i$ to be an identity matrix, our observed sample for n subjects is $(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_n, \mathbf{X}_n)$, with $\mathbf{y}_i$ as the response vector for subject i. The joint data likelihood (without integrating out the random-effects $\mathbf{b}_i$) is given as:

$$l(\Omega; \mathbf{b}, \mathbf{X}, \mathbf{y}) = \sum_{i,j:y_{ij}=0} \log(p_0) + \sum_{i,j:y_{ij}=1} \log(p_1) + \sum_{i,j:y_{ij}\in(0,1)} l_1(\Omega, \mathbf{b}_i, \mathbf{X}_i, \mathbf{y}_i) \tag{4}$$

where

$$\begin{aligned} l_1(\Omega, \mathbf{b}_i, \mathbf{X}_i, \mathbf{y}_i) = \quad & \log \quad [\Gamma(\phi)] - \log\left\{\Gamma\left[\frac{\exp(\mathbf{X}_i^\top\beta + \mathbf{b}_i)}{1+\exp(\mathbf{X}_i^\top\beta + \mathbf{b}_i)}\phi\right]\right\} \\ & - \quad \log\left\{\Gamma\left[\left(1 - \frac{\exp(\mathbf{X}_i^\top\beta + \mathbf{b}_i)}{1+\exp(\mathbf{X}_i^T\beta + \mathbf{b}_i)}\right)\phi\right]\right\} \\ & + \quad \left[\frac{\exp(\mathbf{X}_i^\top\beta + \mathbf{b}_i)}{1+\exp(\mathbf{X}_i^\top\beta + \mathbf{b}_i)}\phi - 1\right]\log(\mathbf{y}_i) \\ & + \quad \left[\left(1 - \frac{\exp(\mathbf{X}_i^\top\beta + \mathbf{b}_i)}{1+\exp(\mathbf{X}_i^\top\beta + \mathbf{b}_i)}\right)\phi - 1\right]\log(1-\mathbf{y}_i) \end{aligned}$$

Although ML estimation of Ω is feasible via maximizing the loglikelihood given in (4) using standard softwares (`SAS`, `R`, etc), challenges remain, such as finding standard errors in multi-parameter problems, and having samples of small to moderate size where the asymptotic theory of MLE may not apply. Recent developments in Markov chain Monte Carlo (MCMC) methods facilitate easy and straightforward implementation of the Bayesian paradigm through conventional software such as `WinBUGS`. The Bayesian approach accommodates full parameter uncertainty through appropriate prior choices supported with proper sensitivity investigations, and provides direct probability statement about a parameter through credible intervals (C.I.) Dunson (2001). Next, we investigate the choice of priors on our model parameters to conduct Bayesian inference.

2.4 Priors, hyperpriors and posterior distributions

We specify practical weakly informative prior opinion on the fixed effects regression parameters β, and non-informative opinion on β_0, the proportions p_0 and p_1, the dispersion parameter ϕ, and

the random effects $\mathbf{b}_i$. Specifically, we assign weakly informative i.i.d Normal$(0, \text{Precision} = 0.01)$ priors on the elements of β, which centers the 'odds-ratio' type inference at 1 with a sufficiently wide 95% interval. For the intercept term β_0, we associate a skeptical flat prior Normal$(0, \text{Precision} = 0.001)$. Priors for p_0 and p_1 are: $p_0 \sim \text{Uniform}(0,1)$ and $p_1 \sim \text{Uniform}(0, 1-p_0)$; $\phi \sim \text{Gamma}(0.1, 0.01)$. Prior on $\mathbf{b}_i$ is zero mean Normal with precision = $1/\sigma_b^2$), where $1/\sigma_b^2 \sim \text{Gamma}(0.01, 0.01)$. Although multivariate specifications (multivariate zero mean vector with inverted-Wishart covariance) are certainly possible, we stick to simple (and independent) choices.

The posterior conclusions will be based on the joint posterior distribution of all the model parameters (conditional on the data), and is obtained combining the log-likelihood given in (4), and the joint prior densities using the Bayes' Theorem:

$$p(\Omega, \mathbf{b}, \sigma_b^2 | \mathbf{X}, \mathbf{y}) \propto l(\Omega; \mathbf{b}, \mathbf{X}, \mathbf{y}) \times \pi_0(\beta) \times \pi_1(\phi) \times \pi_2(p_0) \times \pi_3(p_1) \times \pi_4(\mathbf{b}|\sigma_b^2) \times \pi_5(\sigma_b^2) \quad (5)$$

where $\pi_j(.), j = 0, \ldots, 5$ denote the prior/hyperprior distributions on the model parameters as described above. The relevant the MCMC steps (combination of Gibbs sampling and Metropolis-within-Gibbs) was implemented using the `R2WinBUGS` packageSturtz *et al.* (2005) which connects the `R` with the `WinBUGS` software. After discarding 40000 burn-in samples, we used 40000 more samples (with a spacing of 2) from 2 independent chains with widely dispersed starting values for posterior summaries. Convergence was monitored via MCMC chain histories, autocorrelation and crosscorrelation, density plots, and the Brooks-Gelman-Rubin potential scale reduction factor $\hat{\text{R}}$, all available in the `R coda` library Cowles & Carlin (1996). Associated `R2WinBUGS` code is available on request from the corresponding author.

3 Bayesian model selection and influence diagnostics

3.1 Model selection and assessments

We use the conditional predictive ordinate (CPO) statistic Carlin & Louis (2008) for our model selection, derived from the posterior predictive distribution (ppd). Let $\mathscr{D}$ be the full data, $\mathscr{D}^{(-i)}$ the data with the ith observation deleted, and θ, our parameter vector defined as $\theta = (\beta, p_0, p_1, \phi, \sigma^2)$. We denote the posterior density of $\boldsymbol{\theta}$, given $\mathscr{D}^{(-i)}$ by $\pi(\boldsymbol{\theta}|\mathscr{D}^{(-i)})$. For the i-th observation, the CPO_i can be written as $CPO_i = \int_\Theta f(\mathbf{y}_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathscr{D}^{(-i)})d\boldsymbol{\theta} = \left\{\int_\Theta \frac{\pi(\boldsymbol{\theta}|\mathscr{D})}{f(\mathbf{y}_i|\boldsymbol{\theta})} d\boldsymbol{\theta}\right\}^{-1}$. In absence of a closed form, a Monte Carlo estimate of CPO_i can be obtained using a harmonic-mean approximation Dey *et al.* (1997) as $\widehat{CPO_i} = \left\{\frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{f(\mathbf{y}_i|\boldsymbol{\theta}_q)}\right\}^{-1}$, where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ is a post burn-in sample of size Q from $\pi(\boldsymbol{\theta}|\mathscr{D})$. A summary statistic of the CPO_i's is the log pseudo-marginal likelihood (LPML), defined by $\text{LPML} = \sum_{i=1}^{n} \log(\widehat{CPO_i})$. Larger values of LPML indicates better fit. Because the harmonic-mean identity can be unstable Raftery *et al.* (2007), we consider a more pragmatic route and compute the CPO (and associated LPML) statistics using 500 non-overlapping blocks of the Markov chain, each of size 2000 post-convergence (i.e., after discarding the initial burn-in samples), and report the expected LPML computed over the 500 blocks.

Some other measures, like the DIC, EAIC and EBIC Carlin & Louis (2008) can also be used. Because of the mixture framework in our ZOAB-RE model, we use the DIC_3 Celeux *et al.* (2006) measure, which is an alternative to DIC Spiegelhalter *et al.* (2002). This is defined as

$\text{DIC}_3 = \overline{D(\theta)} + \tau_D$, $\overline{D(\theta)} = -2\text{E}\{\log[f(\mathbf{y}|\theta)]|\mathbf{y}\}$, $f(\mathbf{y}|\theta) = \prod_{i=1}^{n} f(\mathbf{y}_i|\theta)$ is the likelihood function given in (2), $\text{E}\{\log[f(\mathbf{y}|\theta)]|\mathbf{y}\}$ is the posterior expectation of $\log[f(\mathbf{y}|\theta)]$ and τ_D is a measure of the effective number of parameters in the model, given by $\tau_D = \overline{D(\theta)} + 2\log(\text{E}[f(\mathbf{y}|\theta)|\mathbf{y}])$. Thus, we have $\text{DIC}_3 = -4\text{E}\{\log[f(\mathbf{y}|\theta)]|\mathbf{y}\} + 2\log(\text{E}[f(\mathbf{y}|\theta)|\mathbf{y}])$. Let $\theta^{(q)}$ be the MCMC posterior sample generated at the iteration q of the algorithm, $q = 1, \ldots, Q$. The first expectation in this expression can be approximated by $\overline{D} = \frac{1}{Q}\sum_{q=1}^{Q}\sum_{i=1}^{n} \log\left[f(\mathbf{y}_i|\theta^{(q)})\right]$, as recommended by Celeux *et al.* (2006), the second term in the expression can be approximated by $\sum_{i=1}^{n} 2\log \hat{f}(\mathbf{y}_i|\theta)$ with $\hat{f}(\mathbf{y}_i|\theta) = \frac{1}{Q}\sum_{q=1}^{Q} f(\mathbf{y}_i|\theta^{(q)})$. The EAIC and EBIC can be estimated as $\widehat{\text{EAIC}} = -2\overline{D} + 2\nu$ and $\widehat{\text{EBIC}} = -2\overline{D} + \nu\log n$. where ν is the number of parameters in the model, n is the number of observations and $\overline{D}$ defined above. Model selection follows the 'lower is better' law, i.e., the model with the lowest value for these criteria gets selected.

To determine model adequacy after selecting the best model, we use discrepancy measures based on ppd. If $\mathbf{y}_{pr}$ denotes the predictive data vector, then the ppd is given by $p(\mathbf{y}_{pr}|\mathbf{y}) = \int p(\mathbf{y}_{pr}|\theta)p(\theta|\mathbf{y})d\theta$. Samples from the ppd are replicates of the observed model generated data. If the observed value is extreme relative to the reference ppd, there is some concern with respect to model adequacy. We consider three discrepancy measures: (a) the mean statistic $T_1(\mathbf{Y}_j, \theta) = \text{Mean}(\mathbf{Y}_j)$, (b) the minimum statistic $T_2(\mathbf{Y}_j, \theta) = \text{Min}(\mathbf{Y}_j)$, and (c) the maximum statistic $T_3(\mathbf{Y}_j, \theta) = \text{Max}(\mathbf{Y}_j)$. While T_1 assesses the overall fit, T_2 and T_3 are responsible for the tails. Then, the Bayesian p-value Gelman *et al.* (2004) is defined as the number of times $T_i(\mathbf{y}_{pr}, \theta)$ exceeds $T_i(\mathbf{y}, \theta), i = 1, \ldots, 3$ out of L simulated draws, i.e., $p_B^i = \Pr(T_i(\mathbf{y}_{pr}, \theta) \geq T_i(\mathbf{y}, \theta)|y)$. A very large p-value (> 0.95), or a very small (< 0.05) both signals model misspecification.

3.2 Bayesian case influence diagnostics

In this section, we develop some influence diagnostics measures to study the impact of outliers on fixed effects parameter estimates motivated by data perturbation schemes based on case-deletion statistics of Cook & Weisberg (1982). A common way of quantifying influence with and without a given subset of data is to use the q-divergence measures Csisz *et al.* (1967); Weiss (1996) between posterior distributions. Consider a subset I with k elements from the whole dataset with n elements. When the subset I is deleted from the data $\mathbf{y}$, we denote the eliminated data as $\mathbf{y}_I$ and the remaining data as $\mathbf{y}_{(-I)}$. Then, the perturbation function for deletion cases can be written as $p(\theta) = \pi\left(\theta|\mathbf{y}_{(-I)}\right)/\pi\left(\theta|\mathbf{y}\right)$. The q-divergence measure between two arbitrary densities π_1 and π_2 for θ is defined as $d_q(\pi_1, \pi_2) = \int q\left(\frac{\pi_1(\theta)}{\pi_2(\theta)}\right)\pi_2(\theta)d\theta$, where q is a convex function such that $q(1) = 0$. The q-influence of the data $\mathbf{y}_I$ on the posterior distribution of θ, $d_q(I) = d_q(\pi_1, \pi_2)$, is obtained by considering $\pi_1(\theta) = \pi_1(\theta|\mathbf{y}_{(-I)})$ and $\pi_2(\theta) = \pi(\theta|\mathbf{y})$, and can be written as $d_q(I) = \mathbb{E}_{\theta|\mathbf{y}}\{q(p(\theta))\}$, where the expectation is taken with respect to the unperturbed posterior distribution. For various choices of the $q(\cdot)$ function, we have, for example, the Kullback-Leibler (KL) divergence when $q(z) = -\log(z)$, the J-distance (symmetric version of the KL divergence) when $q(z) = (z-1)\log(z)$, and the L_1-distance when $q(z) = |z-1|$.

Note that, $d_q(I)$ defined above precludes itself from quantifying a cut-off point beyond which an observation can be considered influential. Hence, we use the calibration method Peng & Dey (1995), where the probability function of a biased coin is given by $\pi_1(x \mid p) = p^x(1-p)^{1-x}$, with $x = 0, 1$, while that of an unbiased coin is fixed at $\pi_2(x \mid p) = 0.5$. The q-divergence is then $d_q(p) = \frac{q(2p)+q(2(1-p))}{2}$, where $d_q(p)$ increases as p moves away from 0.5, and is symmetric and reaches its minimum value at 0.5. Consequently, if we consider $p \geq 0.90$ (or $p \leq 0.10$) indicative of a strong

bias, then, $d_{L_1}(0.90) = 0.90$. Thus, we can detect an influential observation (using the L_1 distance) when $d_{L_1}(i) \geq 0.80$, $i = 1, \ldots, n$. Similarly, for the KL divergence, we have $d_{KL}(0.90) = 0.51$, and for the J-distance $d_J(0.90) = 0.88$. We consider these cut-off values in this paper.

4 Data analysis and findings

In this section, we apply our proposed ZOAB-RE model to our motivating periodontal data. We start with a short description of the dataset. A study assessing the status and progression of PD among Gullah-speaking African-Americans with Type-2 diabetes was conducted at MUSC via. a detailed questionnaire focusing on demographics, social, medical and dental history. The dataset contain records on 28 teeth (considered full dentition, excluding the 4 third-molars) from 290 subjects, where we focus on quantifying the extent and severity of PD with respect to tooth-types. Our response is: 'The proportion of diseased tooth-sites (with CAL value $\geq$ 3mm), for each of the four tooth types, i.e., incisors, canines, pre-molars and molars, within a subject'. This gives rise to a clustered data framework, where each subject records 4 observations corresponding to the 4 tooth-types. Missing teeth was considered 'missing due to PD' where all sites for that tooth contributed to the diseased category. Subject-level covariables in this dataset include Gender (0=male,1= female), Age of subject at examination (in years), Glycosylated Hemoglobin (HbA1c) status indicator (0=controlled,$<7\%$; 1=uncontrolled,$\geq 7\%$) and smoking status (0=non-smoker,1=smoker). The smokers category comprised of both the current and past smokers. We also considered a tooth-level variable, representing each of the four tooth-types, with 'Canine' as the baseline.

As observed in the density histogram in Figure 1 (Panel a), the data are continuous on the range [0,1]. Due to the presence of a substantial number of 0's (114, 9.83%) and 1's (94, 8.10%), BR might be inappropriate here. Hence, we resort to the ZOAB-RE model, controlling for subject-level clustering. From Equation (3), we now have $g_1(\mu_{ij}) = \psi_{ij} + b_i$, where g_1 is logit, and

$$\psi_{ij} = \beta_0 + \beta_1 \text{Gender}_i + \beta_2 \text{Age}_i + \beta_3 \text{HbA1c}_i + \beta_4 \text{Smoker}_i + \beta_5 \text{Incisor}_{ij} + \beta_6 \text{Premolar}_{ij} + \beta_7 \text{Molar}_{ij}, \tag{6}$$

where β_0 is the intercept, $\beta_1, \ldots, \beta_7$ are the regression parameters, and b_i is the subject-level random effect term. Note that, here the model covariates are regressed over μ_{ij}, which is a function of γ_{ij}, the (conditional) mean $E(Y_{ij}|b_i)$ ($\gamma_{ij} = p_1 + (1 - p_0 - p_1)\mu_{ij}$). Because γ_{ij} is also constrained between 0 and 1, alternatively, one can consider regressing over γ_{ij}. This leads to our choice of 2 competing models:

Model 1 $\text{logit}(\mu_{ij}) = \psi_{ij} + b_i$.

Model 2 $\text{logit}(\gamma_{ij}) = \psi_{ij} + b_i$.

We also fit a non-augmented Beta regression model by transforming the data points y to y' via. the Lemon-squeezer (LS) transformation given by $y' = [y(N-1) + 1/2]/N$ Smithson & Verkuilen (2006), and fit the above regressions to μ with the logit link. This is our (**Model 3**), or the LS model. Although other link functions (such as probit, cloglog, etc) are available, we currently restrict ourselves to the symmetric logit link for this paper whose adequacy is assessed later. Note that, Models 1 and 2 which fits the same dataset can be compared using the model choice criteria described in Section 3.1, but not Model 3 because it considers a transformed dataset. Hence, Model 3 is assessed using plots of empirical cumulative distribution functions (ecdfs) of the fitted values to determine how closely the fits resemble the true data.

In the absence of historical data/experiment, our prior choices follow the specifications described in Section 2.4. Table 1 presents the DIC_3, LPML, EAIC and EBIC values calculated for Models 1 and 2. Notice that, Model 1 (our ZOAB-RE model with regression on μ_{ij}) outperforms

Table 1: Model comparison using DIC_3, LPML, EAIC and EBIC criteria

	Model	
Criterion	1	2
DIC3	2430.98	2513.63
LPML	-624.17	-644.45
EAIC	1234.08	1275.48
EBIC	1294.76	1336.15

Model 2 for all criteria. From Figure 1 (Panel b), it is also clear that the ecdf from the fitted values using Model 1 represent the true data much closely, as compared to Model 3. Considering these, we select Model 1 as our best model, and now proceed to assess its goodness-of-fit using Bayesian p-values. The posterior means of p_B^1, p_B^2 and p_B^3 are respectively 0.273, 0.177 and 0.215, which indicates no overall lack of fit, as well as on the tails.

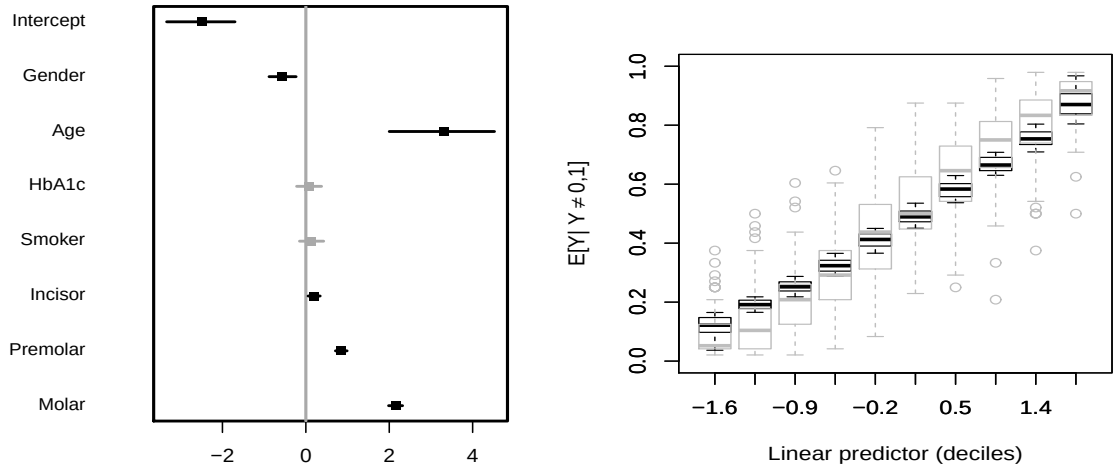

Figure 2: Left panel: posterior median and 95% credible intervals (CI) of parameter estimates from Model 2. CIs that include zero are gray, those that does not include zero are black. Right panel: observed and fitted relationship between the linear predictor ψ_{ij} and the (conditional) non-zero-one mean μ_{ij}. Modeled logit relationships are represented by black box-plots, while the empirical proportions by gray box-plots.

Table 2 presents the posterior parameter estimates obtained from fitting Models 1-3 to the dataset. The 95% CIs are presented only for Model 1 (our best model). For a visual illustration, Figure 2 (left panel) plots the posterior parameter medians, and the 95% CIs of the model parameters for Model 1. The gray intervals in Figure 2 (left panel) contains zero, and refer to non-significant covariates, the black intervals do not include zero and are considered significant at 5% level. The covariates Gender, Age, and the tooth-types appear significant to explain the proportion CAL responses. Conditional on the set of other covariates and REs, parameter interpretation of a covariate can be expressed in terms of its effect directly on μ_{ij}, specifically $\frac{\mu_{ij}}{1-\mu_{ij}}$, and indirectly on the mean of the ZOAB-RE model γ_{ij} which is related to μ_{ij}. Here, μ_{ij} can be defined as the 'condi-

tional expected proportion of diseased sites, conditional on this proportion not being zero or one', and $1-\mu_{ij}$ as the difference to complete disease. Hence, the results in Table 2 can be expressed as the number of times higher/lower the ratio of the conditional expected proportion of diseased sites to the difference to complete disease, is, with every unit increase (for a continuous covariate, such as Age), or a change in category say from 0 to 1 (for a discrete covariate, say Gender). For Age (a strong predictor of PD), this ratio is remarkable high, ($\exp(3.289) = 26.81, 95\%\text{CI} = [7.41, 91.94]$) times higher for every unit increase in Age. For Gender, we can conclude that this ratio is 43% lower for males as compared to females. Although study recruitment design was gender blind, females participated at a higher rate than the males, not unusual for studies on this population Johnson-Spruill *et al.* (2009); Bandyopadhyay *et al.* (2009), and further patient navigator techniques are being developed to achieve better gender balance. The other significant covariates can be interpreted similarly. For example, this ratio is 8.60 times higher for molars as compared to the canines (the baseline), which confirms that the molars which are posteriorly located within the buccal cavity, usually have a higher proportion of diseased tooth-types as compared to the anterior canines. The estimates of p_0 and p_1, 0.099 and 0.082 respectively, closely resemble the true proportion of zeros and ones in the data, and the estimate of ϕ is 7.602, with the $95\%\text{CI} = [6.79, 8.39]$. Note that because γ_{ij} is a monotone function of μ_{ij} for fixed p_0 and p_1 (which is our case), a covariate that serves to increase μ also serves to increase γ. Hence, although our interpretation with μ is somewhat 'indirect' with respect to the response Y, the direction remains the same at the γ level.

Table 2: Posterior parameter (mean) estimates and standard deviations (SD) obtained after fitting Models 1-3 to the periodontal data. The 2.5% and 97.5% percentiles are provided only for Model 1 (our best model).^{s} denotes a significant parameter.

	Model 1				Model 2		Model 3	
Parameter	mean	SD	2.5%	97.5%	mean	SD	mean	SD
Intercept	-2.495^s	0.3986	-3.335	-1.701	-1.771^s	0.293	-3.341^s	0.498
Gender	-0.558^s	0.1626	-0.880	-0.236	-0.353^s	0.113	-0.723^s	0.193
Age	3.289^s	0.6507	2.003	4.521	2.292^s	0.480	4.544^s	0.801
HbA1c	0.084	0.1473	-0.219	0.374	0.037	0.101	0.232	0.169
Smoker	0.126	0.1488	-0.149	0.425	0.095	0.107	0.031	0.172
Incisor	0.198^s	0.0731	0.056	0.341	0.122^s	0.054	0.357^s	0.069
Premolar	0.855^s	0.0704	0.712	0.986	0.585^s	0.055	1.132^s	0.078
Molar	2.151^s	0.0862	1.987	2.317	1.490^s	0.069	2.729^s	0.087
ϕ	7.602	0.4098	6.790	8.386	6.475	0.366	4.655	0.248
p_0	0.099	0.0087	0.084	0.117	0.083	0.008	-	-
p_1	0.082	0.0082	0.066	0.098	0.079	0.008	-	-
σ^2	1.217	0.1287	0.982	1.478	0.574	0.066	1.798	0.182

To investigate the adequacy of the logit link for our regression, we consider an empirical approach via plots of the linear predictor versus the predicted probability Hatfield *et al.* (2012) as depicted in Figure 2. We consider ψ_{ij} from Model 1, and divided it into 10 intervals containing roughly equal number of observations. We again plot the distribution of the inverse-logit transformed linear predictors (denoted by the black box-plots), representing the fitted mean μ_{ij} of the

non-zero-one responses. Next, we overlay the empirical distributions of the observed non-zero-one responses represented by the gray box-plots. From Figure 2, we see no evidence of link misspecification, i.e., the shapes of the fitted and observed trends are similar. As mentioned earlier, one can definitely fit other link functions, however the convenient interpretations in terms of μ_{ij} and γ_{ij} are no longer valid for these fits.

A sensitivity analysis was also conducted on the prior assumptions for the random effects precision parameter $(1/\sigma_b^2)$, fixed effects precision, and for the proportions p_0 and p_1. In particular, we allowed $1/\sigma_b^2 \sim \text{Gamma}(k,k)$, where $k \in \{0.001, 0.1\}$; the normal precision on the fixed effects be 0.1, 0.25 (which reflects an odds-ratio in between e^{-4} to e^4) and 0.001; $p_0 \in \text{Beta}(a,b)$, where $a \in \{1,2,3\}$ and $b \in \{2,3\}$, etc. We checked the sensitivity in the posterior estimates of β by changing one parameter at a time, and refitting Model 1. Although slight changes were observed in parameter estimates and model comparison values, results appear to be robust, and did not change conclusions regarding our best model, inference (and sign) of the fixed-effects, and the influential observations.

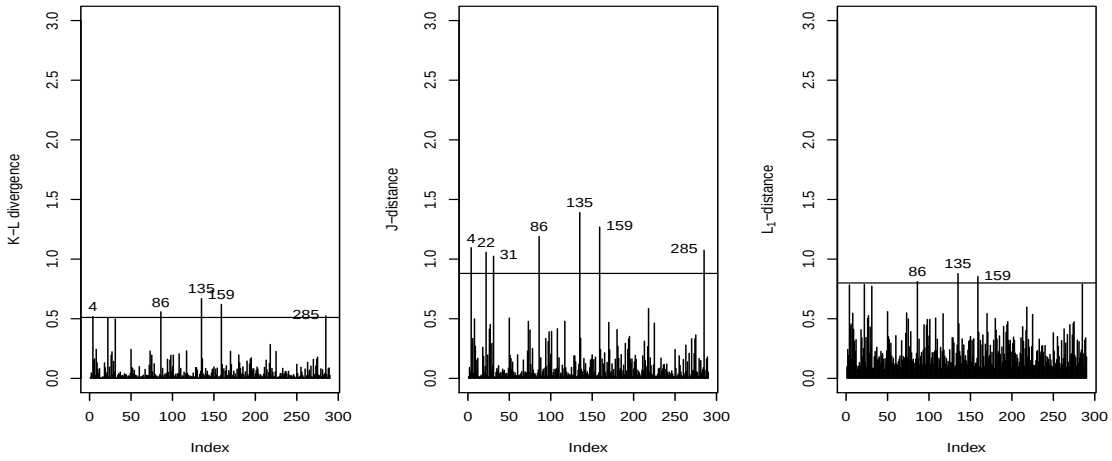

Figure 3: K-L divergence, J and L1 divergence measures for the periodontal dataset.

Table 3: The % change in the posterior (mean) parameter estimates in Model 1, after removing the influential observations. ^{s} indicates parameter significance.

Parameter	#4	#22	#31	#86	#135	#159	#285	All influential obs.
Intercept	-0.080^s	1.122^s	-1.723^s	-1.683^s	-3.246^s	-1.884^s	-4.409^s	-14.749^s
Gender	-3.943^s	5.556^s	-1.613^s	-2.688^s	-1.075^s	-0.896^s	4.480^s	-27.061^s
Age	0.760^s	-3.649^s	1.368^s	1.399^s	4.469^s	-0.517^s	2.919^s	19.033^s
HbA1c	-10.714	-58.333	-44.048	13.095	-21.429	15.476	-26.190	-44.048
Smoker	20.635	44.444	50.794	37.302	6.349	50.794	38.095	137.302
Incisor	-2.525^s	-1.010^s	-3.030^s	-0.505^s	-7.071^s	-6.566^s	0.505^s	-5.051^s
Premolar	0.117^s	1.287^s	0.234^s	0.000^s	-0.117^s	-0.585^s	1.287^s	1.053^s
Molar	-0.046^s	0.046^s	-0.325^s	-0.232^s	-0.325^s	-0.325^s	0.418^s	0.604^s
ϕ	-0.105	-0.368	-0.500	-0.250	-0.605	-0.105	0.158	-1.329
p_0	1.010	0.000	0.000	0.000	0.000	1.010	0.000	2.020
p_1	0.000	0.000	0.000	0.000	0.000	-1.220	0.000	1.220
σ^2	-0.904	-1.726	-1.397	-1.315	-2.301	-2.958	-0.904	-13.394

Finally, to determine the effect of possible influential observations, we computed the q-divergence measures for the best model. Using the cut-offs described in Section 3.2, Figure 3 presents the KL, J and L_1 divergence values for the data points. Specifically, the subjects with id #s $4, 22, 86, 135, 159$ and 285 are considered influential, because they exceed the specified thresholds. To quantify the impact of these observations, we refit the model by removing the data points successively, and also after removing all these influential points. Table 3 records the % change in parameter estimates as compared to the previous posterior estimates present in Table 2. The two binary covariates 'Smoker' and 'HbA1c' are heavily impacted by these observations as compared to the other covariates, the parameter significance and sign of the coefficients remained the same.

5 Simulation Studies

In this section, we conduct a finite sample simulation study to investigate the consequences on parameter estimates after applying the LS transformation to the data in $[0,1]$. Here, the goal is to compare between the mean squared error (MSE), relative bias, and coverage probability for the regression parameter estimates, obtained after fitting (a) our ZOAB-RE model (Model 1) and (b) the LS model (Model 3), for various sample sizes. For the data generation scheme, the location parameter μ_{ij} is generated as: $\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij} + b_i$, with, $b_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$, $j = 1, \ldots, 5$ indicating a cluster of size 5, and various choices of sample sizes $n = 50, 100, 150, 200$. The parameters ϕ, p_0 and p_1 are considered constants, with values $\phi = 2$, $p_0 = 0.1$ and $p_1 = 0.1$. The explanatory variables $x_{ij} = x_i$, are generated as independent draws from a Bernoulli(0.8), and regression parameters and variance components are fixed at: $\beta_0 = 0.5$, $\beta_1 = -0.5$, and $\sigma^2 = 4$. This generates y_{ij}'s in $(0,1)$. The final step is to allocate the 0's, 1's, and the $y_{ij} \in (0,1)$, with probabilities p_0, p_1 and $1 - p_0 - p_1$, which is a result of two successive Bernoulli draws. In the first draw, we set the success probability $(1 - p_0 - p_1)$ for a draw from $y_{ij} \in (0,1)$, else it is either a 0 or 1. Conditional on this draw, we perform another Bernoulli draw with success probability $\frac{p_1}{p_0 + p_1}$ for the occurrence of a 1, else it is a 0.

We simulated 100 such data sets, and fitted the ZOAB-RE and the SL models with similar prior choices as in the data analysis. With our parameter space $\theta = \{\beta_0, \beta_1, \sigma^2, p_0, p_1, \phi\}$, and θ_s an element of θ, we calculate the MSE as $\text{MSE}(\hat{\theta}_s) = \frac{1}{100}\sum_{i=1}^{100}(\hat{\theta}_{is} - \theta_s)^2$, the relative bias as $\text{Relative Bias}(\hat{\theta}_s) = \frac{1}{100}\sum_{i=1}^{100}\left(\frac{\hat{\theta}_{is}}{\theta_s} - 1\right)$, and the 95% coverage probability (CP) as $\text{CP}(\hat{\theta}_s) = \frac{1}{100}\sum_{i=1}^{100} I(\theta_s \in [\hat{\theta}_{s,LCL}, \hat{\theta}_{s,UCL}])$, where I is the indicator function such that θ_s lies in the interval $[\hat{\theta}_{s,LCL}, \hat{\theta}_{s,UCL}]$, with $\hat{\theta}_{s,LCL}$ and $\hat{\theta}_{s,UCL}$ as the estimated lower and upper 95% CIs, respectively. The results from the simulation study for varying sample sizes comparing the ZOAB-RE and the LS models are listed in Table 4, while Figure 4 presents a visual comparison of the models (bold line for the ZOAB-RE model and dashed line for the LS model) for β_0 and β_1. From Table 4, as expected, the absolute values of the relative bias of the parameters are much larger for the SL model as compared to the ZOAB-RE model. Estimates of both p_0 and p_1 show positive relative bias, while it is mostly negative for the other parameters. While comparing MSEs for β_0 (the intercept term) , the LS model seem to perform better than the ZOAB-RE model for low sample sizes, comparable for other sizes, and finally the ZOAB-RE outperforms the LS for $n = 200$. For β_1, the MSEs for the LS model remain lower until it reverses for $n = 200$. For ϕ and σ^2, MSEs are uniformly lower for the ZOAB-RE as compared to the LS model. The CP always remains higher for the ZOAB-RE model as compared to the LS model across all parameters. Interestingly, the CP for ϕ and σ^2 is 0 for the LS model.

Table 4: Relative bias, mean squared error (MSE), and coverage probabilities (CP) of the the parameter estimates after fitting the ZOAB-RE and LS models to simulated data for various sample sizes.

	ZOAB-RE model				SL Model			
Parameter	$n = 50$	$n = 100$	$n = 150$	$n = 200$	$n = 50$	$n = 100$	$n = 150$	$n = 200$
				Relative bias				
β_0	-0.115	-0.260	-0.264	-0.192	-0.613	-0.669	-0.666	-0.642
β_1	-0.076	-0.307	-0.235	-0.189	-0.603	-0.722	-0.678	-0.656
ϕ	0.008	-0.008	-0.001	-0.010	-0.614	-0.648	-0.663	-0.675
p_0	0.052	0.004	0.001	0.011	-	-	-	-
p_1	0.186	0.188	0.146	0.149	-	-	-	-
σ^2	-0.193	-0.226	-0.235	-0.243	-0.893	-0.896	-0.898	-0.899
				MSE				
β_0	0.266	0.147	0.1422	0.065	0.153	0.144	0.146	0.118
β_1	0.348	0.220	0.1598	0.108	0.166	0.173	0.155	0.132
ϕ	0.087	0.028	0.0191	0.014	1.516	1.682	1.760	1.825
p_0	0.001	0.0001	0.0001	0.00001	-	-	-	-
p_1	0.001	0.0001	0.0003	0.0003	-	-	-	-
σ^2	0.958	0.966	0.973	1.053	12.808	12.87	12.91	12.95
				CP				
β_0	98.00	93.00	97.00	98.00	82.00	51.00	53.00	28.00
β_1	97.00	94.00	95.00	96.00	87.00	59.00	61.00	38.00
ϕ	92.00	95.00	96.00	95.00	0.000	0.000	0.000	0.000
p_0	94.00	96.00	98.00	98.00	-	-	-	-
p_1	81.00	76.00	72.00	67.00	-	-	-	-
σ^2	90.00	60.00	49.00	31.00	0.000	0.000	0.000	0.000

In summary, when data is generated from a ZOAB-RE model with responses in $[0,1]$, parameter estimates of the intercept term and fixed effects (β_0 and β_1 here) might not be affected that much when using the LS transform with respect to MSEs, but exhibit poorer performances in terms of relative bias, and coverage probabilities. Hence, the LS transformation might not be adequate even in the presence of a moderate number of 0's and 1's, and further simulation studies (not presented here due to sake of brevity) reveal poorer performances with increase in the proportion of 0's and 1's.

6 Conclusions

Motivated by the classical development of Ospina & Ferrari (2010), we develop a model for clustered responses in $[0,1]$, and apply it to an interesting dataset on PD. We also develop tools for detecting outlying observations using q-divergence measures, and quantify their effect on the posterior estimates of model parameters. Both simulation studies and real data application corroborates seeking an appropriate theoretical model over utilizing some data transformations to the responses for a simpler model. Note that the proposition of Ospina & Ferrari (2010) (without any

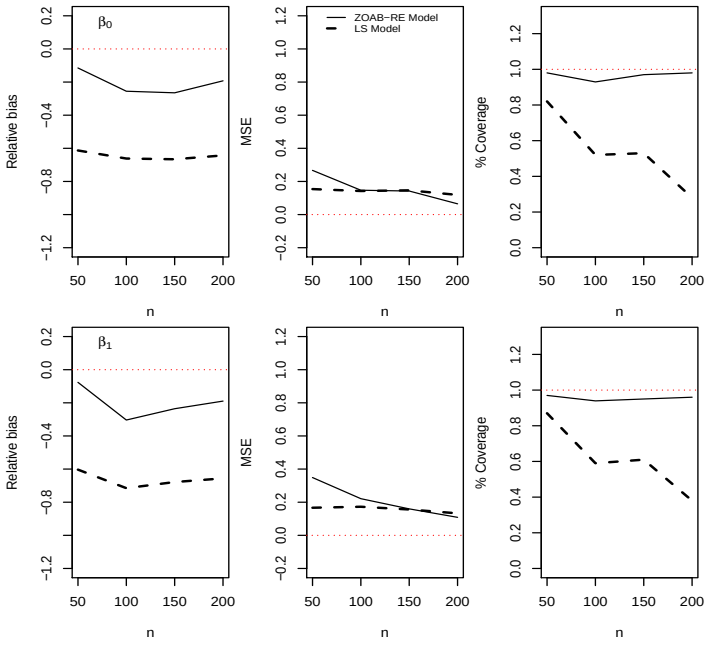

Figure 4: Relative Bias, MSE and coverage for the parameters β_0 and β_1 using ZOAB-RE and SL models.

random effects) is termed 'Inflated Beta distributions'. Typically, for cases of *value-inflation*, such as the zero-inflated counts of Lachenbruch (2002), or the zero-inflated (longitudinal) continuous data as in Ghosh & Albert (2009), inflation occurs when the probability mass of a value exceeds what is allowed by the proposed (underlying) distribution. This is certainly not the case here, and following Hatfield *et al.* (2012), we prefer to call it an 'augmented' model over an 'inflated' model. Our models can be fitted using standard available software packages, such as `R` and `WinBUGS`, with easy access to practitioners in the field.

Our current development treats the parameters p_0, p_1, the probabilities that the response is a 0, or 1, respectively, and ϕ (the dispersion parameter) to be constants, and estimable from the data. This assumption is primarily to seek a parsimonious model to address the homogeneity of the subjects (all Type-2 diabetic African-Americans), and also our focus on quantifying covariate-response relationships for responses which are still 'at risk' of developing PD (i.e., $Y \in (0,1)$). The cases $Y = 0$ and $Y = 1$ represent extreme (disease-free and completely diseased) situations, and we tackle this by augmenting these extremities to the original BR proposition, and connect the model covariates to μ_{ij}, the conditional expected proportion of diseased sites in $(0,1)$. Although regressing over γ_{ij} (the conditional expectation of the true augmented Beta response, conditional over random effects) was not worthwhile as in Model 2 for this dataset, it might certainly be important in a different application. Also, other applications might consider the probabilities p_0 and p_1 varying with subject and the clustered units, i.e., p_{0ij} and p_{1ij}, and estimating them from the data (via priors), or regressing over model covariates. It is also of interest to investigate presence of thick/heavy tails in the underlying ZOAB-RE proposition, and model the random effect term b_i using a robust t-density over a Normal density, as in Figueroa-ZúñIga *et al.* (2012). For our dataset, the results were very similar using a t-density, and hence we did not consider that route.

Our current analysis considers a 'clustered' cross-sectional periodontal proportion data. Often,

these study subjects might be randomized to dental treatments, and subsequent longitudinal follow-ups, leading to a clustered-longitudinal proportion data, where one might be interested to estimate the profiles (both overall, and subject level) in the proportion of diseased surfaces of the four tooth-types with time. Our ZOAB-RE can certainly be extended to such situations with proper consideration to the GLMM random effects specifications. Other propositions available in the literature on modeling clustered (or longitudinal) proportion responses include simplex mixed-effects models Qiu *et al.* (2008), robust transformation modelsSong & Tan (2000); Zhang *et al.* (2009), etc. How these models compare with ours, and ways to adapt these to proportion responses in $[0, 1]$ are immediate components of future research, and will be considered elsewhere.

Acknowledgements

The authors thank the Center for Oral Health Research at MUSC for providing the motivating data, and Prof. Elizabeth Slate for interesting insights on clinical interpretations. Galvis acknowledges support from CAPES/CNPq, Brasil. Bandyopadhyay acknowledges support from grants R03DE020114 and R03DE021762 from the US National Institutes of Health. Lachos was supported by grants 305054/2011-2 from CNPq-Brazil and 2011/17400-6 from FAPESP, Brazil.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall, Ltd.

Bandyopadhyay, D., Reich, B. J. & Slate, E. H. (2009). Bayesian modeling of multivariate spatial binary data with applications to dental caries. *Statistics in medicine*, **28**(28), 3492–3508.

Branscum, A., Johnson, W. & Thurmond, M. (2007). Bayesian beta regression: Applications to household expenditure data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics*, **49**(3), 287–301.

Carlin, B. & Louis, T. (2008). *Bayesian Methods for Data Analysis (Texts in Statistical Science)*. Chapman and Hall/CRC, New York,.

Celeux, G., Forbes, F., Robert, C. P. & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, **1**(4), 651–673.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B,*, **48**, 133–169.

Cook, R. D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall/CRC, Boca Raton, FL.

Cowles, M. K. & Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**(434), 883–904.

Csisz, I. et al. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, **2**, 299–318.

Dey, D. K., Chen, M. H. & Chang, H. (1997). Bayesian approach for the nonlinear random effects models. *Biometrics*, **53**, 1239–1252.

Dunson, D. (2001). Commentary: practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, **153**(12), 1222.

Fernandes, J., Salinas, C., London, S., Wiegand, R., Hill, E., Slate, E., Grewal, J., Werner, P., Sanders, J. & Lopes-Virella, M. (2006). Prevalence of periodontal disease in gullah african american diabetics. *J Dent Res*, **85**, 997.

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

Figueroa-ZúñIga, J. I., Arellano-Valle, R. B. & Ferrari, S. L. (2012). Mixed beta regression: A bayesian perspective. *Computational Statistics & Data Analysis*.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian data analysis*. Chapman & Hall/CRC.

Ghosh, P. & Albert, P. S. (2009). A bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial. *Computational statistics & data analysis*, **53**(3), 699–706.

Hatfield, L. A., Boye, M. E., Hackshaw, M. D. & Carlin, B. P. (2012). Multilevel bayesian models for survival times and longitudinal patient-reported outcomes with many zeros. *J. Am. Stat. Assoc.*, **107**, 875–885.

Johnson, N., Kotz, S. & Balakrishnan, N. (1994). *Continuous Univariate Distributions, Vol. 2*. New York: John Wiley & Sons. ISBN 978-0-471-58494-0.

Johnson-Spruill, I., Hammond, P., Davis, B., McGee, Z. & Louden, D. (2009). Health of gullah families in south carolina with type 2 diabetes diabetes self-management analysis from project sugar. *The Diabetes Educator*, **35**(1), 117–123.

Kieschnick, R. & McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling*, **3**(3), 193–213.

Lachenbruch, P. A. (2002). Analysis of data with excess zeros. *Statistical Methods in Medical Research*, **11**(4), 297–302.

Lachos, V. H., Bandyopadhyay, D. & Dey, D. K. (2011). Linear and nonlinear mixed-effects models for censored hiv viral loads using normal/independent distributions. *Biometrics*, **67**(4), 1594–1604.

Ospina, R. & Ferrari, S. (2010). Inflated beta distributions. *Statistical Papers*, **51**(1), 111–126.

Peng, F. & Dey, D. K. (1995). Bayesian analysis of outlier problems using divergence measures. *The Canadian Journal of Statistics*, **23**, 199–213.

Qiu, Z., SONG, P. X.-K. & Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics*, **35**(4), 577–596.

Raftery, A., Newton, M., Satagopan, J. & Krivitsky, P. (2007). *Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion)*, volume 8, pages 1–45. Oxford University Press.

Simas, A., Barreto-Souza, W. & Rocha, A. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis*, **54**(2), 348–366.

Smithson, M. & Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, **11**(1), 54.

Song, P. X.-K. & Tan, M. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics*, **56**(2), 496–502.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society-Series B*, **64**(4), 583–639.

Sturtz, S., Ligges, U. & Gelman, A. (2005). R2winbugs: A package for running winbugs from r. *Journal of Statistical Software*, **12**(3), 1–16.

Verkuilen, J. & Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution. *Journal of Educational and Behavioral Statistics*, **37**(1), 82–113.

Weiss, R. (1996). An approach to bayesian sensitivity analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(4), 739–750.

Zeileis, A., Cribari-Neto, F. & Grün, B. (2010). Beta regression in r. *Journal of Statistical Software*, **34**(2), 1–24.

Zhang, P., Qiu, Z., Fu, Y. & Song, P. X.-K. (2009). Robust transformation mixed-effects models for longitudinal continuous proportional data. *Canadian Journal of Statistics*, **37**(2), 266–281.