UNICAMP

DIANA MILENA GALVIS SOTO

BAYESIAN ANALYSIS OF REGRESSION MODELS FOR PROPORTIONAL DATA IN THE PRESENCE OF ZEROS AND ONES

ANALISE BAYESIANA DE MODELOS DE REGRESSÃO PARA DADOS DE PROPORÇÕES NA PRESENÇA DE ZEROS E UNS

CAMPINAS
2014

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA
E COMPUTAÇÃO CIENTÍFICA

DIANA MILENA GALVIS SOTO

BAYESIAN ANALYSIS OF REGRESSION MODELS FOR PROPORTIONAL DATA IN THE PRESENCE OF ZEROS AND ONES

ANALISE BAYESIANA DE MODELOS DE REGRESSÃO PARA DADOS DE PROPORÇÕES NA PRESENÇA DE ZEROS E UNS

Thesis presented to the Institute of Mathematics, Statistics and Scientific Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Statistics

Tese apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutora em Estatística

Orientador: Víctor Hugo Lachos Dávila

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE
DEFENDIDA PELA ALUNA DIANA MILENA GALVIS SOTO,
E ORIENTADA PELO PROF. DR VÍCTOR HUGO LACHOS DÁVILA

Assinatura do Orientador

____

CAMPINAS
2014

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

G139b

Galvis Soto, Diana Milena, 1978-
Bayesian analysis of regression models for proportional data in the presence of zeros and ones / Diana Milena Galvis Soto. – Campinas, SP : [s.n.], 2014.

Orientador: Víctor Hugo Lachos Dávila.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Inferência bayesiana. 2. Modelos mistos. 3. Dados de proporções. 4. Dados agrupados. 5. Doença periodontal. I. Lachos Dávila, Víctor Hugo,1973-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Análise bayesiana de modelos de regressão para dados de proporções na presença de zeros e uns
Palavras-chave em inglês:
Bayesian Inference
Mixed models
Proportional data
Clustered data
Periodontal disease
Área de concentração: Estatística
Titulação: Doutora em Estatística
Banca examinadora:
Víctor Hugo Lachos Dávila [Orientador]
Nancy Lopes Garcia
Silvia Lopes de Paula Ferrari
Marcos Oliveira Prates
Mário de Castro Andrade Filho
Data de defesa: 25-09-2014
Programa de Pós-Graduação: Estatística

Tese de Doutorado defendida em 25 de setembro de 2014 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). VÍCTOR HUGO LACHOS DÁVILA

Nancy L. Garcia

Prof(a). Dr(a). NANCY LOPES GARCIA

Prof(a). Dr(a). SILVIA LOPES DE PAULA FERRARI

Prof(a). Dr(a). MARCOS OLIVEIRA PRATES

Mário de Castro Andrade Filho

Prof(a). Dr(a). MÁRIO DE CASTRO ANDRADE FILHO

Abstract

Continuous data in the unit interval $(0, 1)$ represent, generally, proportions, rates or indices. However, zeros and/or ones values can be observed, representing absence or total presence of a carachteristic of interest. In that case, regression models that analyze the effect of covariates such as beta, beta rectangular or simplex are not appropiate. In order to deal with this type of situations, an alternative is to add the zero and/or one values to the support of these models. In this thesis and based on these models, we propose the mixed regression models for proportional data augmented by zero and one, which allow analyze the effect of covariates into the probabilities of observing absence or total presence of the interest characteristic, besides of being possivel to deal with correlated responses. Estimation of parameters can follow via maximum likelihood or through MCMC algorithms. We follow the Bayesian approach, which presents some advantages when it is compared with classical inference because it allows to estimate the parameters even in small size sample. In addition, in this approach, the implementation is straightforward and can be done using software as `openBUGS` or `winBUGS`. Based on the marginal likelihood it is possible to calculate selection model criteria as well as q-divergence measures used to detect outlier observations.

Keywords: Bayesian inference, mixed models, proportional data, clustered data, periodontal disease.

Resumo

Dados no intervalo (0,1) geralmente representam proporções, taxas ou índices. Porém, é possível observar situações práticas onde as proporções sejam zero e/ou um, representando ausência ou presença total da característica de interesse. Nesses casos, os modelos que analisam o efeito de covariáveis, tais como a regressão beta, beta retangular e simplex não são convenientes. Com o intuito de abordar este tipo de situações, considera-se como alternativa aumentar os valores zero e/ou um ao suporte das distribuições previamente mencionadas. Nesta tese, são propostos modelos de regressão de efeitos mistos para dados de proporções aumentados de zeros e uns, os quais permitem analisar o efeito de covariáveis sobre a probabilidade de observar ausência ou presença total da característica de interesse, assim como avaliar modelos com respostas correlacionadas. A estimação dos parâmetros de interesse pode ser via máxima verossimilhança ou métodos Monte Carlo via Cadeias de Markov

(MCMC). Nesta tese, será adotado o enfoque Bayesiano, o qual apresenta algumas vantagens em relação à inferência clássica, pois não depende da teoria assintótica e os códigos são de fácil implementação, através de softwares como openBUGS e winBUGS. Baseados na distribuição marginal, é possível calcular critérios de seleção de modelos e medidas Bayesianas de divergência q, utilizadas para detectar observações discrepantes.

Palavras-chave: Inferência Bayesiana, modelos mistos, dados de proporções, dados agrupados, doença periodontal.

Contents

A Dios, mi esposo, mi madre y mis lindos hijos Juan Pablo y Ana Sofía.

Acknowledgements

First, I am thankful to our awesome god for his constant presence in my life.

I would like to express my deepest gratitude to CAPES/CNPQ - IEL Nacional Brazil by the economic support during my PhD studies. Thanks by the support.

I would like to thank my husband and my mother by their support, love and understanding, without them this PhD would have been just my biggest dream and not my greatest academic achievement. Also, I would like thank my children by their support in periods of absence and also by the immense amount of love that they always give me.

I would like to express my deepest gratitude to my advisor and friend Víctor Hugo Lachos Dávila by his guidance, patience and by his confidence in myself.

Also, I would like to thank Professores Dipankar and Mauricio Castro by sharing with me their knowledge. I am very grateful for all that I have learned.

I also appreciate all the assistance that I received from my teachers Marina, Nancy, Víctor and Ronaldo for sharing with me the knowledge they have acquired during their lives.

I am very thankful to my friends Diego, Marcos and José Alejandro, by their support and company during my first and most difficult year away from my family.

I also would like to thank Maria Cecilia Baranauskas from the Institute of Computation at UNICAMP, advisor of my husband Julian, for allow me to join my family here in Brazil.

There are many other people who I am very thankful, People from the secretary of pos-graduate students at IMECC, to Mrs Zefa, who always offered me his affection and friendship, and also to all my friends of IMECC, who I not will list because would be huge.

I want to thank the Center for Oral Health Research at MUSC for providing the motivating data for this thesis.

List of Figures

List of Tables

Chapter 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 and 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 and Tan, 2000). Some of the strategies pointed out in the statistical literature to analyze this type of data are based on regression models combined with a particular data transformation such as the *logit* transformation. However, the use of nonlinear transformations may hinder the interpretation of the regression parameters. This situation can be overcome by considering probability distributions with double-bounded support, such as the beta, simplex (Barndorff-Nielsen and Jørgensen, 1991), and beta rectangular distributions (Hahn, 2008), which can be parameterized in terms of their mean. Based on these models, regression models were proposed.

The beta regression (BR) reparameterizes the associated beta parameters, connecting the response to the data covariates through suitable link functions (Ferrari and Cribari-Neto, 2004). Yet, the beta density does not accommodate tail-area events, or flexibility in variance specifications (Bayes et al., 2012). To accommodate this, the BRe density Hahn (2008), and associated regression modelsBayes et al. (2012) were considered under a Bayesian framework. Note, the BRe regression includes the (constant dispersion) BRFerrari and Cribari-Neto (2004), and the variable dispersion BRSmithson and Verkuilen (2006) as special cases. The simplex regressionSong and Tan (2000) is based on the simplex distribution from the dispersion family (Jørgensen, 1997), assumes constant dispersion, and uses extended generalized estimating equations for inference connecting the mean to the covariates via the logit link. Subsequently, frameworks with heterogenous dispersion (Song et al., 2004), and for mixed-effects models (Qiu et al., 2008) were explored. Yet, their potential were limited to proportion responses with support in $(0, 1)$.

The methodology developed in this thesis is motivated from a study conducted at the Medical University of South Carolina (MUSC) via a detailed questionnaire focusing on demographics, social, medical and dental history. In this study, was assessed the status and progression of periodontal disease (PrD) among Gullah-speaking African-Americans with Type-2 diabetes (Fernandes et al., 2006). The dataset contain measurements from Clinical Attachment Level (CAL), obtained as the distance between the soft tissue in relation to the cemento-enamel junction (see Figure 1.1), on six different sites of each one of 28 teeth (considered full dentition, excluding the 4 third-molars). In this study were observed 290 subjects, recording proportion of diseased tooth-sites. The proportion is calculated as the number of sites with the disease divided by the number of sites. This number depends on the type of tooth, for example, there are 48 sites for molar, premolar and incisive, but 24 sites for canine. A site is said to be diseased if the value of CAL is $\geq$ 3mm. Hence, this clustered data framework has 4 observations (corresponding to the 4 tooth-types) for each subject. If a tooth is missing, it was considered 'missing due to PrD', where all sites for that tooth contributed to the diseased category. Note that in this case, the response lies in the closed interval [0,1]; where 0 and 1 represent completely disease free and highly diseased cases, respectively.

Subject-level covariables in the dataset include gender (0=male, 1= female), age of subject at examination (in years, ranging from 26 to 87 years), glycosylated hemoglobin (HbA1c) status indicator (0=controlled,$<$ 7%; 1=uncontrolled,$\geq$ 7%) and smoking status (0=non-smoker,1=smoker). We also considered a tooth-level variable representing each of the four tooth types, with 'canine' as the baseline.

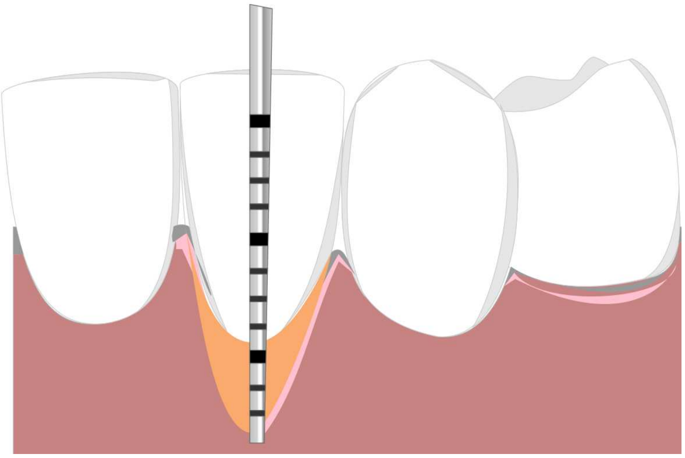

Figure 1.1: Clinical attachment level.

The underlying statistical question here 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, premolars and molars)' with the covariables.

In this thesis are presented mixed regression models for proportional data in the presence of zeros and ones as an alternative when the response is a vector with correlated components in $[0, 1]$. In this case, the estimation of the parameters, model selection and influence diagnostics follows a Bayesian approach.

1.1 Preliminaries

1.1.1 Bayesian model selection and influence diagnostics

Model selection and assessments

There is a variety of methods for selecting the model that best fit a dataset. In this thesis will be used the log pseudo-marginal likelihood (LPML), the observed information criterion DIC$_3$, the expected Akaike information criterion (EAIC) and the expected Bayesian information criterion EBIC.

The LPML (Geisser and Eddy, 1979) is a summary statistic of the conditional predictive ordinate (CPO) statistic and is defined by LPML $= \sum\limits_{i=1}^{n} \log(\widehat{CPO}_i)$, where $\widehat{CPO}_i$ can be obtained using a harmonic-mean approximation Dey et al. (1997) as $\widehat{CPO}_i = \left\{ \frac{1}{Q} \sum\limits_{q=1}^{Q} \frac{1}{f(\mathbf{y}_i|\boldsymbol{\theta}^{(q)})} \right\}^{-1}$ and $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(Q)}$ is a post burn-in sample of size Q from the posterior distribution of $\boldsymbol{\theta}$ and $f(\mathbf{y}_i|\boldsymbol{\theta}^{(q)})$ is the marginal distribution of Y. Larger values of LPML indicates better fit.

Some other measures, like the DIC, EAIC and EBIC Carlin and Louis (2008) can also be used. Because of the mixture framework in our models, we use the DIC$_3$ Celeux et al. (2006) measure, which is an alternative to DIC Spiegelhalter et al. (2002). This is defined as DIC$_3 = \overline{D(\boldsymbol{\theta})} + \tau_D$, $\overline{D(\boldsymbol{\theta})} = -2\mathrm{E}\{\log[f(\mathbf{y}|\boldsymbol{\theta})]|\mathbf{y}\}$, $f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\theta})$, $\mathrm{E}\{\log[f(\mathbf{y}|\boldsymbol{\theta})]|\mathbf{y}\}$ is the posterior expectation of $\log[f(\mathbf{y}|\boldsymbol{\theta})]$ and τ_D is a measure of the effective number of parameters in the model, given by $\tau_D = \overline{D(\boldsymbol{\theta})} + 2\log(\mathrm{E}[f(\mathbf{y}|\boldsymbol{\theta})|\mathbf{y}])$. Thus, we have DIC$_3 = -4\mathrm{E}\{\log[f(\mathbf{y}|\boldsymbol{\theta})]|\mathbf{y}\} + 2\log(\mathrm{E}[f(\mathbf{y}|\boldsymbol{\theta})|\mathbf{y}])$. The first expectation in this expression can be approximated by $\overline{D} = \frac{1}{Q}\sum_{q=1}^{Q}\sum_{i=1}^{n} \log\left[f(\mathbf{y}_i|\boldsymbol{\theta}^{(q)})\right]$, as recommended by Celeux et al. (2006), the second term in the expression can be approximated by $\sum_{i=1}^{n} 2\log \hat{f}(\mathbf{y}_i|\boldsymbol{\theta})$ with $\hat{f}(\mathbf{y}_i|\boldsymbol{\theta}) = \frac{1}{Q}\sum_{q=1}^{Q} f(\mathbf{y}_i|\boldsymbol{\theta}^{(q)})$. The EAIC and EBIC can be estimated as $\widehat{\mathrm{EAIC}} = -2\overline{D} + 2\nu$ and $\widehat{\mathrm{EBIC}} = -2\overline{D} + \nu \log n$, 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.

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 and 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 $p(\boldsymbol{\theta}) = \pi\left(\boldsymbol{\theta}|\mathbf{y}_{(-I)}\right)/\pi\left(\boldsymbol{\theta}|\mathbf{y}\right)$. The q-divergence measure between two arbitrary densities π_1 and π_2 for $\boldsymbol{\theta}$ is defined as $d_q(\pi_1, \pi_2) = \int q\left(\frac{\pi_1(\boldsymbol{\theta})}{\pi_2(\boldsymbol{\theta})}\right)\pi_2(\boldsymbol{\theta})d\boldsymbol{\theta}$, 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 $\boldsymbol{\theta}$, $d_q(I) = d_q(\pi_1, \pi_2)$, is obtained by considering $\pi_1(\boldsymbol{\theta}) = \pi_1(\boldsymbol{\theta}|\mathbf{y}_{(-I)})$ and $\pi_2(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$, and can be written as $d_q(I) = \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y}}\left\{q(p(\boldsymbol{\theta}))\right\}$, 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 and Dey (1995), in that work, the q-divergence is given by $d_q(p) = \frac{q(2p)+q(2(1-p))}{2}$, where $d_q(p)$ increases as p moves away from 0.5, and is symmetric and reaches its minimum value at 0.5. It is possible consider $p \geq 0.90$ (or $p \leq 0.10$), however, in this thesis will be used $p = 0.95$. Thus, we can detect an influential observation (using the L_1 distance) when $d_{L_1}(i) \geq 0.90$, $i = 1, \ldots, n$, for the KL divergence, we have $d_{KL}(0.95) = 0.83$, and for the J-distance $d_J(0.95) = 1.32$.

1.2 Organization of Thesis

This thesis is divided into five chapters and four appendices. The second chapter is an already published paper and the fourth chapter is a paper that has been recently accepted for publication. In the fifth chapter are presented the conclusions and the plan for future research. Next, I describe the results of chapters second to fourth.

- Chapter 2: Augmented mixed beta regression models for periodontal proportion data, published paper in *Statistics in Medicine (2014).*

  - **Description:** In this chapter was proposed the Bayesian analysis of the zero and one augmented mixed beta regression (ZOAB-RE) model for clustered responses in $[0, 1]$, and applied it to an interesting PrD dataset. Through this model it was possible to identifying covariates that are significant to explain disease-free, progressing with disease, and completely diseased tooth types. We also developed tools for outlier detection using q-divergence measures, and quantified their effect on the posterior estimates of the model parameters. Both simulation studies and real data application justify seeking an appropriate theoretical model over utilizing ad hoc data transformations for proportion data

- Chapter 3: Augmented mixed models for clustered proportion data using the simplex distribution.

  - **Description:** In this chapter was proposed a Bayesian random effect model based on the simplex distribution for modeling data in the interval $[0, 1]$ called zero and one mixed simplex regression (ZOAS-RE) model. The versatility of this class to model correlated data in the interval $[0, 1]$ has not been explored elsewhere, and this is our major contribution. Simulation studies reveal good consistency properties of the Bayesian estimates when compared with the ZOAS-RE regression counterpart, as well as, high performance of the model selection techniques to pick the appropriately fitted model

- Chapter 4: Augmented mixed models for clustered proportion data. Accepted paper in *Statistical Methods in Medical Research (2014)*.
  - **Description:** In this chapter, it was proposed a class of (parametric) augmented proportion distribution models. Particular cases of this family are the beta, beta rectangular and simplex distributions. Based on this distributions were proposed the regression models under a Bayesian framework, and demonstrate its application to the PrD dataset. Also, these regression models were compared using the PrD dataset and simulation studies. The results allow conclude that the ZOAS-RE model fits better to the PrD than the ZOAB-RE and zero and one augmented beta rectangular (ZOABr-RE) models. It was also concluded via simulation studies.

Chapter 2

Augmented mixed beta regression models for periodontal proportion data

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 heavily diseased subjects in any study, the proportion values can lie in the interval $[0, 1]$. While Beta regression can be adapted to assess covariate effects in these situations, its 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 ad hoc alternatives and provide quantitative insight into assessing the true covariate effects on the proportion responses.

2.1 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 and 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 and Tan, 2000). With fidelity to the usual Gaussian assumptions for model errors, one might here be tempted 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 logit-transformed proportion responses, can provide a computationally convenient framework, but 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 and Cribari-Neto (2004) can accomplish direct modeling of covariates under a generalized linear model (GLM) specification, leading to easy interpretation. The beta density (Johnson et al., 1994) is extremely flexible, and can take on a variety of shapes to account for non-normality and skewness in proportion data. The BR model considers a specific re-parameterization 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 (Fernandes et al., 2006), where the clinical attachment level (or, CAL), a clinical marker of periodontal disease (PrD), is measured at each of the 6 sites of a subject's tooth. The underlying statistical question here 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, premolars and molars)' with the covariables. Figure 1 (left panel) plots the raw (unadjusted) density histogram of the proportion responses aggregated 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 (ad hoc) re-scaling (Smithson and Verkuilen, 2006) of the data from $[0, 1]$ to the interval $(0, 1)$, various limitations are observed working on a transformed scale (Lachos et al., 2011). These re-scalings might provide a nice working solution for small proportions of 0's and 1's, but sensitivity towards parameter estimation can be considerable with higher proportions. This inefficiency is only aggravated due to the presence of additional clustering (tooth within mouth/subject) in the data, as in our case. Hence, from a practical perspective, there is a need to seek an appropriate theoretical model that avoids data transformations, yet is capable of handling the challenges the data present. 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, which can accommodate the subject-level clustering.

There have been various specifications of the BR model. The BR model of Ferrari and 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 by linking the covariates to the dispersion parameter via the variable dispersion BR model by Smithson and Verkuilen (2006). Very recently, Verkuilen and Smithson (2012) used Gauss-Hermite quadrature to calculate ML estimates and a Gibbs sampler for Bayesian estimation in the context of BR models for correlated proportion data. Also, Figueroa-Zúniga et al. (2013) presents a Bayesian approach to the correlated BR model through Gibbs samplers, and uses the deviance information criterion (DIC) (Spiegelhalter et al., 2002), expected-AIC (EAIC) and expected-BIC (EBIC) for model selection. However, to the best of our knowledge, there are no studies that utilize a Bayesian paradigm to model clustered (correlated) proportion data where the proportions lie in the interval [0,1]. Our proposition 'augments' point masses

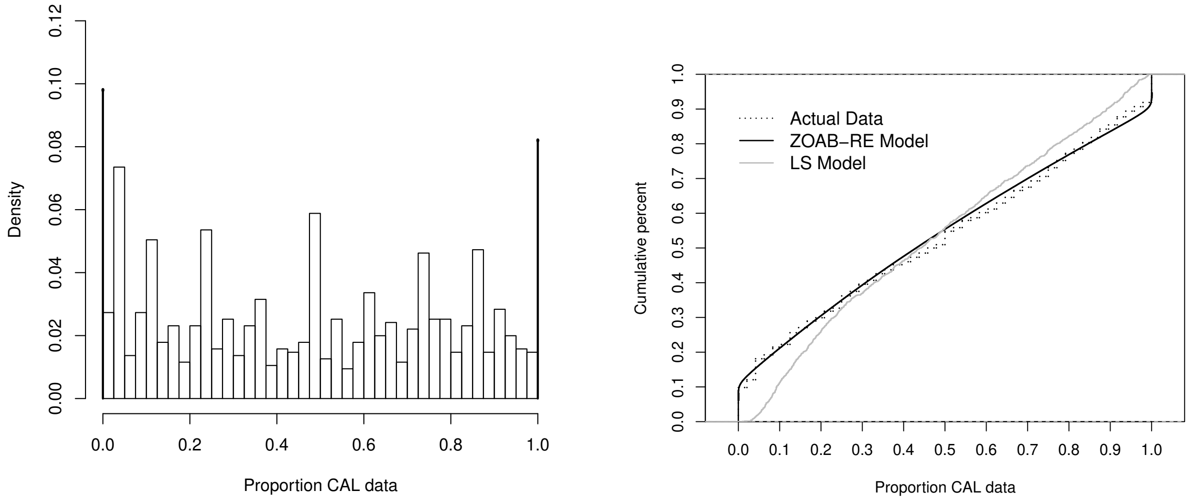

Figure 2.1: The left panel plots the (raw) density histogram, aggregated over subjects and tooth-types for the PrD data. The 'pins' at the extremes represent the proportion of zeros (9.8%) and ones (8.1%). The right panel presents the empirical cumulative distribution function of the real data, and that obtained after fitting the ZOAB-RE (Model 1) and the LS model (Model 3).

at zero and one to a continuous (beta) density that does not include zero and one in its support, similar in spirit to 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 `OpenBUGS` (Thomas et al., 2006).

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 applies the proposed ZOAB-RE model to the motivating data and uses Bayesian model selection to select the best model. It also summarizes and discusses the estimation of the fixed effects, other model parameters and outlier detections. Section 4 presents simulation studies to assess finite sample performance of our model with another competing transformation-based model under model misspecification, and also to study the efficiency of the influence diagnostic measures to detect outliers. Conclusions and future developments appear in Section 6.

2.2 Statistical Model and Bayesian Inference

2.2.1 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 μ and a precision parameter ϕ is given by

$$f(y|\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \; 0<y<1, \; 0<\mu<1, \; \phi>0, \quad (2.2.1)$$

where $\Gamma(\cdot)$ denotes the gamma function, $E(Y)=\mu$, and $\text{Var}(Y) = \dfrac{\mu(1-\mu)}{1+\phi}$. Therefore, for a fixed value of the mean μ, higher values of ϕ leads to a reduction of $\text{Var}(Y)$, and conversely. If Y has pdf as in (2.2.1), we write $Y \sim \text{beta}(\mu\phi;(1-\mu)\phi)$. Next, to connect the covariate vector $\mathbf{x}_i$ to the random sample $Y_1,\ldots,Y_n$ of Y, we use a suitable link function g_1 that maps the mean interval (0,1) onto the real line. This is given as $g_1(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ 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 constant (Ferrari and Cribari-Neto, 2004), or regressed onto the covariates (Smithson and Verkuilen, 2006) via another link function h_1, such that $h_1(\phi_i) = \mathbf{z}_i^T\boldsymbol{\alpha}$, where $\mathbf{z}_i$ is a covariate vector (not necessarily similar to $\mathbf{x}_i$), and $\boldsymbol{\alpha}$ is the corresponding vector of regression parameters. Similar to $\mathbf{x}_i$, $\mathbf{z}_i$ also accommodates an intercept. Both g_1 and h_1 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 $\Phi(\cdot)$ is the standard normal density, the complementary log-log function $g_1(\mu_i) = \log\{-\log(1-\mu_i)\}$ among others, and for h_1, the log function $h_1(\phi_i) = \log(\phi)$, the square-root function $h_1(\phi_i) = \sqrt{\phi_i}$, and the identity function $h_1(\phi_i) = \phi_i$ (with special attention to the positivity of the estimates) (Simas et al., 2010). Estimation follows via either the (classical) maximum likelihood (ML) route (Ferrari and Cribari-Neto, 2004) or through Gauss-Hermite quadratures (Smithson and Verkuilen, 2006) available in the `betareg` library in `R` (Zeileis et al., 2010), or Bayesian (Branscum et al., 2007) through Gibbs sampling.

2.2.2 Zero-and-one augmented beta random effects model

The BR model described above only applies to observations that are independent, and moreover it is suitable only for responses lying in $(0,1)$. However, for our PrD 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 data clustering. 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 $Y_i \in [0,1]$. Thus, $Y \sim \text{ZOAB}(p_{0_i}, p_{1_i}, \mu_i, \phi)$, if the density of Y_i, $i=1,\ldots,n$, follows

$$f(y_i|p_{0_i},p_{1_i},\mu_i,\phi) = \begin{cases} p_{0_i} & \text{if } y_i = 0, \\ p_{1_i} & \text{if } y_i = 1, \\ (1-p_{0_i}-p_{1_i})f(y_i|\mu_i,\phi) & \text{if } y_i \in (0,1), \end{cases} \quad (2.2.2)$$

where $p_{0_i} \geq 0$ denotes $P(Y_i = 0)$, $p_{1_i} \geq 0$ denotes $P(Y_i = 1)$, $0 \leq p_{0_i} + p_{1_i} \leq 1$ and $f(y_i|\mu_i, \phi)$ is given in (2.2.1). The mean and variance of Y_i are given by

$$\begin{aligned} E[Y_i] &= (1 - p_{0_i} - p_{1_i})\mu_i + p_{1_i} \\ \text{and} \\ \text{Var}(Y_i) &= p_{1_i}(1 - p_{1_i}) + (1 - p_{0_i} - p_{1_i}) \left[\frac{\mu_i(1-\mu_i)}{1+\phi} + (p_{0_i} + p_{1_i})\mu_i^2 - 2\mu_i p_{1_i} \right]. \end{aligned}$$

For clustered data, the ZOAB-RE model is 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 onto a suitably transformed μ_{ij}, p_{0ij} and p_{1ij}, such that

$$g_1(E[\mathbf{Y}_i|\mathbf{b}_i]) = g_1(\boldsymbol{\mu}_i) = \mathbf{X}_i^\top \boldsymbol{\beta} + \mathbf{Z}_i^\top \mathbf{b}_i, \tag{2.2.3}$$

$$g_2(\boldsymbol{p_0}_i) = \boldsymbol{W_0}_i^\top \boldsymbol{\psi}, \tag{2.2.4}$$

$$g_3(\boldsymbol{p_1}_i) = \boldsymbol{W_1}_i^\top \boldsymbol{\rho}, \tag{2.2.5}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})$, $\boldsymbol{p_0}_i = (p_{0i1}, \ldots, p_{0in_i})$, $\boldsymbol{p_1}_i = (p_{1i1}, \ldots, p_{1in_i})$; $\mathbf{X}_i$, $\boldsymbol{W_0}_i$ and $\boldsymbol{W_1}_i$ are design matrices of dimension $p \times n_i$, $r \times n_i$ and $s \times n_i$, corresponding to the vectors of fixed effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_r)^\top$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_s)^\top$, respectively, 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, g_2 and g_3 here remain the same as for g_1 in Subsection 2.2.1. For the sake of interpretation, we prefer to use the logit link. Note that in our model development, the dispersion parameter ϕ is chosen as constant and the regressions onto $\boldsymbol{p_0}_i$ and $\boldsymbol{p_1}_i$ are free of REs to avoid over-parameterization. However, it is certainly possible to regress ϕ onto covariates through an appropriate link function (say, log). Also, p_{0ij} and p_{1ij} can be treated as constants across all sample units. To this end, we define our ZOAB-RE model as $Y_{ij} \sim$ ZOAB-RE$(p_{0_{ij}}, p_{1_{ij}}, \mu_{ij}, \phi)$ $i = 1, \ldots, n,\ j = 1, \ldots, n_i$.

2.2.3 Likelihood function

Let $\boldsymbol{\Omega} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\rho}, \phi)$ denote the parameter vector in this ZOAB-RE model. The primary goal here is to estimate $\boldsymbol{\Omega}$, and to derive inference on $\boldsymbol{\beta}$ adjusting for the effects of clustering. Our observed sample for n subjects is $(\mathbf{y}_1, \mathbf{X}_1, \mathbf{Z}_1, \boldsymbol{W_0}_1, \boldsymbol{W_1}_1), \ldots, (\mathbf{y}_n, \mathbf{X}_n, \mathbf{Z}_n, \boldsymbol{W_0}_n, \boldsymbol{W_1}_n)$, 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:

$$L(\boldsymbol{\Omega}|\mathbf{b}, \mathbf{y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{W_0}, \boldsymbol{W_1}) = \prod_{i=1}^{n} L_i(\boldsymbol{\Omega}|\mathbf{b}_i, \mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{W_0}_i, \boldsymbol{W_1}_i), \tag{2.2.6}$$

where

$$L_i(\boldsymbol{\Omega}|\mathbf{b}_i, \mathbf{y}_i, \mathbf{X}_i, \mathbf{Z}_i, \boldsymbol{W_0}_i, \boldsymbol{W_1}_i) \;=\; \odot \left[\boldsymbol{p_0}_i^\top \mathbf{D}_{0i} + \boldsymbol{p_1}_i^\top \mathbf{D}_{1i} + (1 - \boldsymbol{p_0}_i - \boldsymbol{p_1}_i)^\top (\mathbf{I}_{n_i} - \mathbf{D}_{0i} - \mathbf{D}_{1i}) \mathbf{B}_i \right]^\top,$$

$\odot \mathbf{A}_i$ indicates the product of the elements of $\mathbf{A}_i$, $\boldsymbol{p_0}_i = (p_{0i1}, \ldots, p_{0in_i})^\top$ with $p_{0ij} = \dfrac{\exp(\boldsymbol{W_0}_{ij}^\top \boldsymbol{\psi})}{1 + \exp(\boldsymbol{W_0}_{ij}^\top \boldsymbol{\psi})}$, $\boldsymbol{p_1}_i = (p_{1i1}, \ldots, p_{1in_i})^\top$, with $p_{1ij} = \dfrac{\exp(\boldsymbol{W_1}_{ij}^\top \boldsymbol{\rho})}{1 + \exp(\boldsymbol{W_1}_{ij}^\top \boldsymbol{\rho})}$, $\mathbf{D}_{ki}$ is a diagonal matrix of dimension $n_i \times n_i$ whose j-th element of the diagonal is the indicator function $I_{\{y_{ij}=k\}}$, $k = 0, 1$, $j = 1, \ldots, n_i$, $\mathbf{I}_{n_i}$ is the identity matrix with dimension $n_i \times n_i$ and $\mathbf{B}_i$ is a diagonal matrix of dimension $n_i \times n_i$ whose j-th element of the diagonal is $\frac{\Gamma(\phi)}{\Gamma(\mu_{ij}\phi)\Gamma((1-\mu_{ij})\phi)} y_{ij}^{\mu_{ij}\phi-1}(1-y_{ij})^{(1-\mu_{ij})\phi-1}$ and $\mu_{ij} = \dfrac{\exp(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\top}\mathbf{b}_i)}{1+\exp(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\top}\mathbf{b}_i)}$, $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ correspond to the j-th column of the matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$, respectively.

Although one can certainly pursue a classical estimation route using maximum likelihood methods following Ospina and Ferrari (2010), a Bayesian treatment of our model has not been considered earlier in the literature. Recent developments in Markov chain Monte Carlo (MCMC) methods facilitate easy and straightforward implementation of the Bayesian paradigm through conventional software such as `OpenBUGS`. Hence, we consider a Bayesian estimation framework which can accommodate full parameter uncertainty through appropriate prior choices supported by proper sensitivity investigations. This framework can provide a direct probability statement about a parameter through credible intervals (C.I.) (Dunson, 2001). Next, we investigate the choice of priors for our model parameters to conduct Bayesian inference.

2.2.4 Prior and posterior distributions

We specify practical weakly informative prior opinion on the fixed effects regression parameters $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, $\boldsymbol{\rho}$, ϕ (dispersion parameter) and the random effects $\mathbf{b}_i$. Specifically, we assign i.i.d Normal$(0, \text{precision} = 0.01)$ priors on the elements of $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ and $\boldsymbol{\rho}$, which centers the 'odds-ratio' type inference at 1 with a sufficiently wide 95% interval. Priors for $\phi \sim$ Gamma$(0.1, 0.01)$, and $\mathbf{b}_i$ are Normal with zero mean and precision $= 1/\sigma_b^2)$, where $\sigma_b \sim$ Unif$(0, 100)$ (Gelman, 2006). Although multivariate specifications (multivariate zero mean vector with inverted-Wishart covariance) are certainly possible, we stick to simple (and independent) choices. For cases where p_0 and p_1 are considered constants across all subjects, we allocate the Dirichlet prior with hyperparameter $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ for the probability vector $(p_0, p_1, 1 - p_0 - p_1)$, where $\alpha_s \sim$ Gamma$(1, 0.001)$, $s = 1, 2, 3$.

The posterior conclusions are based on the joint posterior distribution of all the model parameters (conditional on the data), and obtained by combining the likelihood given in (2.2.6), and the joint prior densities using the Bayes' Theorem:

$$p(\boldsymbol{\theta}, \mathbf{b}|\mathbf{y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{W_0}, \boldsymbol{W_1}) \propto L(\boldsymbol{\Omega}|\mathbf{b}, \mathbf{y}, \mathbf{X}, \mathbf{Z}, \boldsymbol{W_0}, \boldsymbol{W_1}) \times \pi_0(\boldsymbol{\beta}) \times \pi_1(\boldsymbol{\psi}) \times \pi_2(\boldsymbol{\rho}) \times \pi_3(\phi) \times \pi_4(\mathbf{b}|\sigma_b) \times \pi_5(\sigma_b), \tag{2.2.7}$$

where $\boldsymbol{\theta} = (\boldsymbol{\Omega}, \sigma_b^2)^{\top}$, $\pi_j(.), j = 0, \ldots, 5$ denote the prior/hyperprior distributions on the model parameters as described above. The relevant MCMC steps (combination of Gibbs and Metropolis-within-Gibbs sampling) was implemented using the `BRugs` package (Ligges et al., 2009) which connects the `R` language with the `OpenBUGS` software. After discarding 50000 burn-in samples, we used 50000 more samples (with spacing of 10) from two 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 and Carlin, 1996). Associated `BRugs` code is available on request from the corresponding author.

2.2.5 Bayesian model selection and influence diagnostics

We use the conditional predictive ordinate (CPO) statistic (Carlin and Louis, 2008) for our model selection derived from the posterior predictive distribution (ppd). A summary statistic obtained from the CPO is the log pseudo-marginal likelihood (LPML) (Carlin and Louis, 2008). Larger values of LPML indicate better fit. Because the harmonic-mean identity used in the CPO computation 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 deviance information criteria (DIC), expected AIC (EAIC) and expected BIC (EBIC) (Carlin and 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 as an alternative to the DIC (Spiegelhalter et al., 2002). 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 apply the Bayesian p-value (Gelman et al., 2004) which utilizes some discrepancy measures based on ppd. Samples from the ppd (denoted by $\mathbf{y}_{pr}$) are replicates of the observed model generated data $\mathbf{y}$, hence there is some signal of model inadequacy if the observed value is extreme relative to the reference ppd. Because of the clustered nature of our data, we consider the sum statistic $T(\mathbf{y}, \boldsymbol{\theta}) = \text{sum}(\mathbf{y})$ as our discrepancy measure. Then, the Bayesian p-value p_B is calculated as the number of times $T(\mathbf{y}_{pr}, \boldsymbol{\theta})$ exceeds $T(\mathbf{y}, \boldsymbol{\theta})$ out of L simulated draws, i.e., $p_B = \Pr(T(\mathbf{y}_{pr}, \boldsymbol{\theta}) \geq T(\mathbf{y}, \boldsymbol{\theta}) | \mathbf{y})$. A very large p-value (> 0.95), or a very small one (< 0.05) signals model misspecification.

In addition, some influence diagnostic measures are developed to study the impact of outliers on fixed effects parameter estimates caused by data perturbation schemes based on case-deletion statistics (Cook and Weisberg, 1982), and the q-divergence measures (Csisz et al., 1967; Weiss, 1996; Lachos et al., 2013) between posterior distributions. We use three choices of these divergences, namely, the Kullback-Leibler (KL) divergence, the J-distance (symmetric version of the KL divergence), and the L_1-distance. We use the calibration method (Peng and Dey, 1995) to obtain the cut-off values as 0.90, 0.83 and 1.32 for the L_1, KL and J-distances, respectively.

2.3 Data analysis and findings

In this section, we apply our proposed ZOAB-RE model to the PrD data. We start with a short description of the dataset. A study assessing the status and progression of PrD among Gullah-speaking African-Americans with Type-2 diabetes (Fernandes et al., 2006) was conducted at the Medical University of South Carolina (MUSC) via a detailed questionnaire focusing on demographics, social, medical and dental history. CAL was recorded at each of the 6 tooth-sites per tooth for 28 teeth (considered full dentition, excluding the 4 third-molars). With 290 subjects, we focus on quantifying the extent and severity of PrD for the tooth-types (4 canines and 8 each of incisors, pre-molars and molars). Our response variable is: 'Proportion of diseased tooth-sites (with CAL value $\geq$ 3mm) for each of the four tooth types'. This gives rise to a clustered data framework where each subject records 4 observations corresponding to the 4 tooth-types. Missing teeth were considered 'missing due to PrD', 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, ranging from 26 to 87 years), glycosylated hemoglobin (HbA1c) status indicator (0=controlled,$< 7\%$; 1=uncontrolled,$\geq 7\%$) and smoking status (0=non-smoker,1=smoker). The smoker category is 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 2.1 (Panel left), the data are continuous in the range [0,1]. Due to the presence of a substantial number of 0's (114, 9.8%) and 1's (94, 8.1%), BR might be inappropriate here. Hence, we resort to the ZOAB-RE model, controlling for subject-level clustering. From Equation (2.2.3), we now have $\boldsymbol{\eta}_i = g_1(\boldsymbol{\mu}_i) = \mathbf{X}_i^\top \boldsymbol{\beta} + \mathbf{b}_i$, with g_1 the logit link, $\boldsymbol{\beta}^\top = (\beta_0, \ldots, \beta_7)$, with β_0 the intercept and $\beta_1, \ldots, \beta_7$ the regression parameters, $\mathbf{X}_i^\top = (1, \text{Gender}_i, \text{Age}_i, \text{HbA1c}_i, \text{Smoker}_i, \text{Incisor}_i, \text{Premolar}_i, \text{Molar}_i)$, and b_i is the subject-level random effect term. To improve convergence of the sampler, we standardized 'Age' by subtracting its mean and dividing by its standard deviation. Note that, here the model covariates are regressed onto μ_{ij}, p_{0ij} and p_{1ij}, but it is also possible to consider p_0 and p_1 constants across all subjects. This leads to our choice of two competing models:

	Model	
Criterion	1	2
DIC$_3$	**993.0**	1243.5
LPML	$-$**500.5**	-623.7
EAIC	**992.7**	1231.0
EBIC	**1124.2**	1286.6

Table 2.1: Model comparison using DIC$_3$, LPML, EAIC and EBIC criteria.

Model 1: $\text{logit}(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i$, $\text{logit}(\boldsymbol{p}_{\mathbf{0}i}) = \boldsymbol{W}_{\mathbf{0}i}^\top \boldsymbol{\psi}$ and $\text{logit}(\boldsymbol{p}_{\mathbf{1}i}) = \boldsymbol{W}_{\mathbf{0}i}^\top \boldsymbol{\rho}$, with $\boldsymbol{W}_{\mathbf{0}i}^\top = \boldsymbol{W}_{\mathbf{1}i}^\top = \mathbf{X}_i^\top$.

Model 2: $\text{logit}(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i\ \boldsymbol{p}_{\mathbf{0}i} = p_0$ and $\boldsymbol{p}_{\mathbf{1}i} = p_1$.

We also fit a non-augmented BR 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 and Verkuilen, 2006), where N is the total number of observations, and fit the above regressions to $\boldsymbol{\mu}_i$ 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 whose adequacy is assessed later. Note that Models 1 and 2 which fit the same dataset can be compared using the model choice criteria described in Subsection 2.2.5, but not Model 3 since 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.

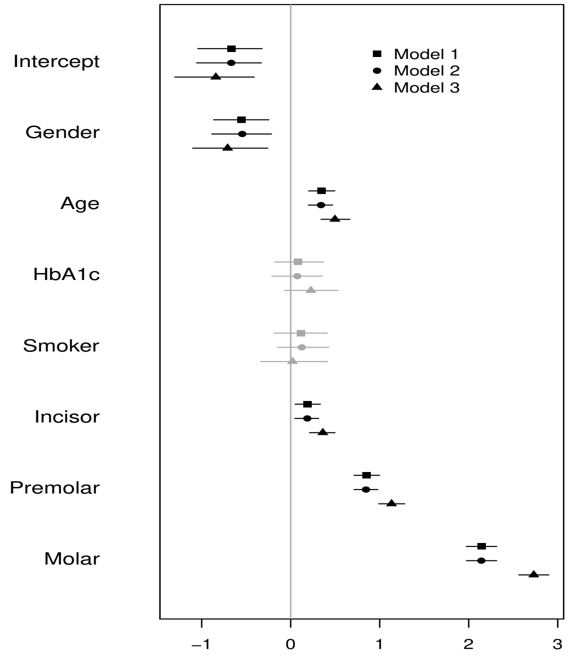

Figure 2.2: Posterior mean and 95% CIs of parameter estimates from Models 1-3. CIs that include zero are gray, those that does not include zero are black.

Table 2.2: The values are the number of times higher/lower the ratio of the conditional 'expected proportion of diseased sites' (denoted by μ_{ij}) is, to the 'expected remaining proportion to complete disease' (denoted by 1 - μ_{ij}), conditional on this proportion not being zero or one, with one unit increase in the covariates.

Parameter	Model 1	Model 2	Model 3
Intercept	0.5	0.5	0.4
Gender	0.6	0.6	0.5
Age	1.4	1.4	1.6
HbA1c	1.1	1.1	1.3
Smoker	1.1	1.1	1.0
Incisor	1.2	1.2	1.4
Premolar	2.3	2.3	3.1
Molar	8.5	8.5	15.3

In the absence of historical data/experiment, our prior choices follow the specifications described in Section 2.2.4. Table 2.1 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}, p_{0ij} and p_{1ij}) outperforms Model 2 for all criteria. From Figure 2.1 (right panel), it is also clear that the ecdf from the fitted values using Model 1 represent the true data more closely than Model 3. Considering these, we select Model 1 as our best model. With respect to goodness-of-fit assessment, $p_B = 0.798$, which indicates no overall lack of fit. Figure 2.2 plots the posterior parameter means and the 95% credible intervals (CIs) for the regression onto $\boldsymbol{\mu}$ for Models 1-3. The gray intervals in Figure 2.2 contain zero (the non-significant covariates), while the black intervals do not include zero (the significant ones at 5% level). The covariates gender, age, and the tooth types (incisor, premolar and molar) significantly explain the proportion responses. Conditional on the set of other covariates and REs, parameter interpretation can be expressed in terms of the corresponding covariate effect directly on μ_{ij}, specifically the ratio $\frac{\mu_{ij}}{1-\mu_{ij}}$. Here, μ_{ij} is the 'expected proportion of diseased sites', and $1 - \mu_{ij}$ is the complement, i.e., the 'expected remaining proportion to being completely diseased', both conditional on Y_{ij} not being zero or one. Hence, the results in Table 2.2 can be expressed as the number of times the ratio is higher/lower 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 example, this ratio for age (a strong predictor of PrD) is $(1.4, 95\%\text{CI} = [1.2, 1.6])$. For gender, we conclude that this ratio is 40% lower for males as compared to females. Although study recruitment design was gender blind, females

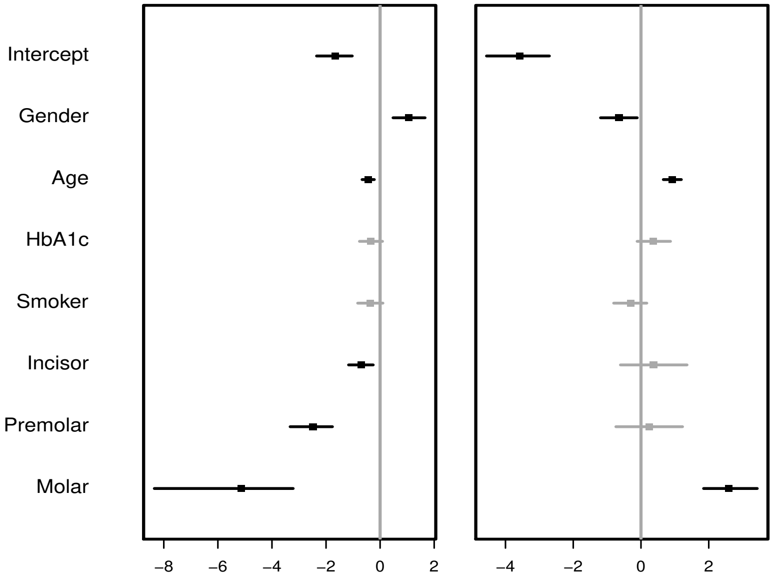

Figure 2.3: Posterior mean and 95% CI of parameter estimates for p_{0ij} (left panel) and p_{1ij} (right panel) from Model 1. CIs that include zero are gray, those that does not include zero are black.

Table 2.3: The values corresponding to p_{0ij} represent odds of having a 'disease free' versus 'diseased' tooth-type, while those for p_{1ij} denote odds of 'completely diseased' versus 'diseased and disease-free' tooth types.

Parameter	p_{0ij}	p_{1ij}
Intercept	0.2	0.03
Gender	2.9	0.5
Age	0.6	2.5
HbA1c	0.7	1.4
Smoker	0.7	0.7
Incisor	0.5	1.4
Premolar	0.08	1.3
Molar	0.005	13.3

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 8.5 times higher for the posteriorly located molars as compared to anteriorly placed canines (the baseline).

The mean estimates (standard deviations) of ϕ for the Models 1, 2 and 3 are 7.6 (0.42), 7.6 (0.43) and 4.6 (0.26), and those of σ_b^2 are 1.2 (0.13), 1.2 (0.13) and 1.8 (0.18), respectively. Based on these and from Table 2.2, we conclude there is little difference between the Models 1 and 2 with respect to the estimates of $\boldsymbol{\beta}$, ϕ and σ_b^2. The main advantage of Model 1 is that it identifies significant covariates related to free PrD and completely diseased tooth types, which is not available in Model 2. However, the estimates of premolar, molar, ϕ and σ_b^2 obtained from Model 3 are greater than those obtained from Models 1 and 2, with the highest difference being for molar. Interestingly, the estimates of ϕ (σ_b^2) from Model 3 are smaller (greater) than those from Models 1 and 2, implying that augmenting leads to a lower (estimated) variance of Y than the transformation-based Model 3.

Figure 2.3 plots the posterior parameter means and the 95% CIs of the parameters used to model $\boldsymbol{p_0}$ (left panel) and $\boldsymbol{p_1}$ (right panel) for Model 1. Gender, age and type of tooth significantly explain free of PrD, while gender, age and molar significantly explain the completely diseased category. Table 2.3 presents the number of times higher/lower of the odds for free of PrD (second column) and completely diseased (third column). For example, the odds of a tooth type free of PrD are 2.9 times greater for men than for women, while the odds of a completely diseased molar are about 13 times that that of a (baseline) Canine. Interestingly, the odds of a completely diseased tooth type are 2.5 times higher for a unit

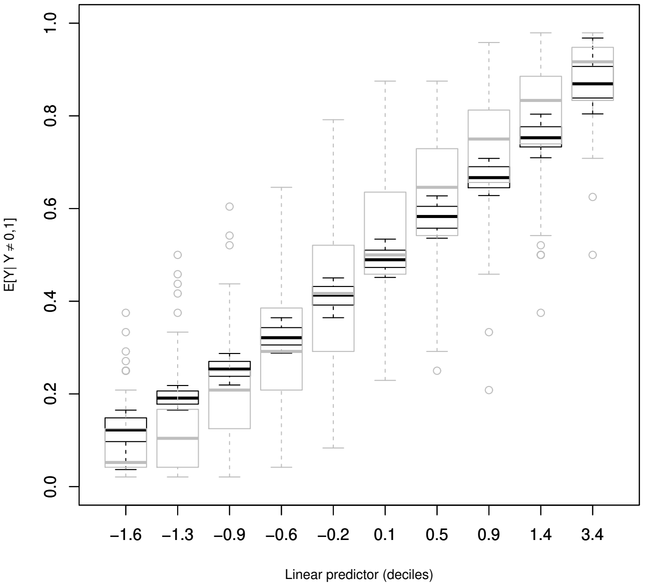

Figure 2.4: Observed and fitted relationship between the linear predictor η_{ij} and the (conditional) non-zero-one mean μ_{ij}. Modeled logit relationships are represented by black box-plots, while the empirical proportions by gray box-plots.

increase in age. Interpretation for the other parameters is similar.

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 2.4. We consider η_{ij} from Model 1, and divide it into 10 intervals containing roughly an equal number of observations. We 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 2.4, we observe 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, but the convenient interpretations in terms of μ_{ij} are no longer valid for these fits.

We also conducted a sensitivity analysis on the prior assumptions for the random effects precision $(1/\sigma_b^2)$ and the fixed effects precision parameter. In particular, we allowed $\sigma_b \sim \text{Uniform}(0, k)$, where $k \in \{10, 50\}$ and also the typical Inverse-gamma choice for the precision $1/\sigma_b^2 \sim \text{Gamma}(k, k)$, where $k \in \{0.001, 0.1\}$. We also chose the normal precision on the fixed effects to be 0.1, 0.25 (which reflects an odds-ratio in between e^{-4} to e^{4}) and 0.001. We checked the sensitivity in the posterior estimates of $\boldsymbol{\beta}$ by changing one parameter at a time, and refitting Model 1. Although slight changes were observed in parameter estimates and model comparison values, the results appeared to be robust, and did not change our conclusions regarding the best model, inference (and sign) of the fixed-effects, and the influential observations.

Finally, to determine the effect of possible influential observations, we computed the q-

divergence measures for Model 1. In particular, the subjects with id # 135, 159, 174 and 285 were considered influential because the values of the L_1, KL and J-distances exceeded the specified thresholds. The subjects $135, 159$ and 285 have higher proportion responses for all tooth types (with $Y_{ij} \geq 0.75$) for than the corresponding mean proportions across all subjects. On the contrary, subject 174 is free of PrD ($Y_{ij} = 0$) across all tooth types. To quantify the impact of these observations on the covariate effects, we refit the model by first removing these subjects successively, and then as a whole. Compared to other covariates, the estimate of molar for the regression onto p_{0ij} was impacted substantially. A minor impact on smoker for regression onto p_{0ij} was also observed when all influential observations were removed. Overall, parameter significance and signs of the coefficients remained the same. Henceforth, we assert to use the estimates obtained from fitting Model 1 to the full data without removing these subjects.

2.4 Simulation studies

In this section, we conduct two simulation studies. For the first, we plan to investigate the consequences on the (regression) parameter estimation under model misspecification via mean squared error (MSE), relative bias (RB), and coverage probability for the (a) ZOAB-RE model (Model 1), and (b) the LS model (Model 3) for varying sample sizes. In the second, we evaluate the efficiency of the q-divergence measures to detect atypical observations in the ZOAB-RE model.

Simulation 1: We generate $T_{ij} \sim \text{Normal}(\mu_{ij}, 1)$, where $i = 1, \ldots, n$ (the number of subjects), $j = 1, \ldots, 5$ (indicating cluster of size 5 for each subject), with location parameter μ_{ij} modeled as $\mu_{ij} = \beta_0 + \beta_1 x_{ij} + b_i$, and $b_i \sim N(0, \sigma^2)$. Then, $y_{ij} = \frac{\exp(T_{ij})}{1+\exp(T_{ij})}$. We choose various sample sizes $n = 50, 100, 150, 200$. The explanatory variables x_{ij} are generated as independent draws from a Uniform$(0, 1)$, and regression parameters and variance components are fixed at $\beta_0 = -0.5$, $\beta_1 = 0.5$, and $\sigma^2 = 2$. This generates data from a logit-normal model with $y_{ij} \in (0, 1)$. Next, we can have two sets of p_0 and p_1; namely, Case (a): $p_0 = 0.01,\ p_1 = 0.01$ and Case (b): $p_0 = 0.1,\ p_1 = 0.08$ (representative of the real data). 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 achieved via multinomial sampling. To keep the simulation design simple, we do not consider the regressions onto p_0 and p_1.

In the first simulation study, we simulated 500 such data sets and fitted the ZOAB-RE and the LS models with similar prior choices as in the data analysis. With our parameter vector $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_b^2, p_0, p_1)$, and θ_s an element of $\boldsymbol{\theta}$, we calculate the MSE as $\text{MSE}(\hat{\theta}_s) = \frac{1}{500}\sum_{i=1}^{500}(\hat{\theta}_{is} - \theta_s)^2$, the relative bias as Relative Bias$(\hat{\theta}_s) = \frac{1}{500}\sum_{i=1}^{500}\left(\frac{\hat{\theta}_{is}}{\theta_s} - 1\right)$, and the 95% coverage probability (CP) as $\text{CP}(\hat{\theta}_s) = \frac{1}{500}\sum_{i=1}^{500} I(\theta_s \in [\hat{\theta}_{s,LCL}, \hat{\theta}_{s,UCL}])$, 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% limitis of the CIs, respectively. Figure 2.5 presents a visual comparison of the parameters β_0 and β_1 for varying sample sizes and proportions p_0 and p_1, where the black and gray lines represent the ZOAB-RE model and the LS model, respectively.

As expected, both panels of Figure 2.5 reveal that the absolute values of RB for both β_0

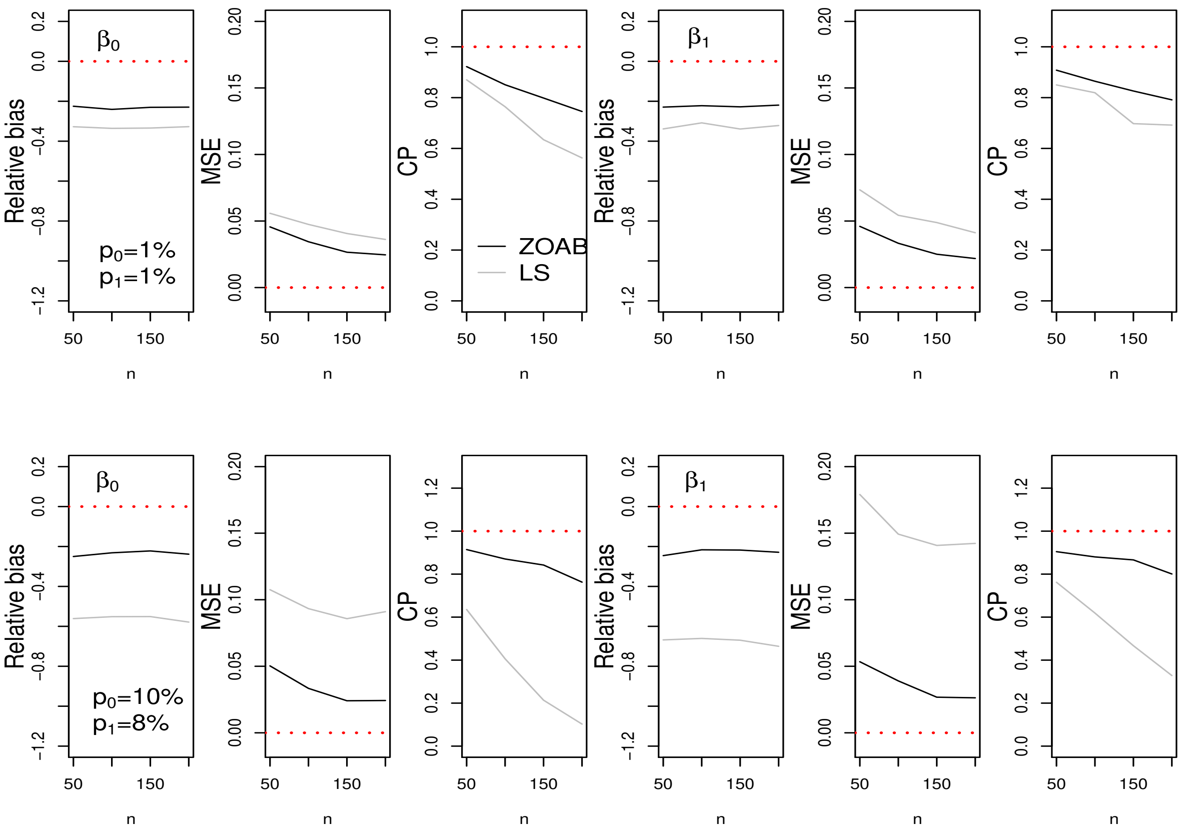

Figure 2.5: Relative Bias, MSE and CP of β_0 and β_1 after fitting the ZOAB-RE (black line) and LS (gray line) models, with $p_0 = p_1 = 1\%$ (upper panel) and $p_0 = 10\%$,$p_1 = 8\%$ (lower panel).

and β_1 are much larger for the LS model than the ZOAB-RE model, with the RB increasing with increasing p_0 and p_1 (Case b). We observe similar behavior for MSE and CP, i.e., both the parameters from the ZOAB-RE model are estimated with lower MSE and higher CP when compared to the corresponding ones from the LS model, with the performance of the LS model getting worse with increasing proportions of extreme values. Clearly, when data are generated from a misspecified (augmented logit-normal) model, the LS model seems to produce a considerable impact on the regression parameter estimates as compared to the more robust ZOAB-RE model. For the sake of brevity, the MSE, RB and CP for the other parameters (p_0, p_1, σ_b^2) are not presented here, but we discuss the results. The proportions p_0 and p_1 are estimated with positive RB. Interestingly, for σ_b^2, the RB remains negative for all cases, with the absolute value of the RB increasing with increasing sample size mainly for the LS model. This might occur because the LS transformation induces lower variability in the data leading to an underestimated σ_b^2 and RB. With this increase in RB, the 95% CI does not include the true value of σ_b^2, and hence the CP is mostly 0 for higher n (150 and 200) for both models in Case (a), and also for all sample sizes for the LS model in Case (b). We conclude that under model misspecification, applying the LS transformation may not be adequate even for a moderate number of 0's and 1's, with the performance deteriorating further as the proportion of extremes increases.

Simulation 2: Here, we simulated one data set with 100 subjects using the same data generation scheme as in Simulation 1. We perturb the response vector for ID #20 via $\mathbf{y}_{20} = \mathbf{y}_{20} + 2\text{sd}(\mathbf{y}_{20})$, where sd stands for standard deviation. If an element of the perturbed vector was greater than 1, we assigned 1 there. Figure 2.6 presents the q-divergence measures, both without perturbation (upper panel) and with perturbation (lower panel). We conclude from here that the divergence measures can correctly detect the influential (perturbed) observations.

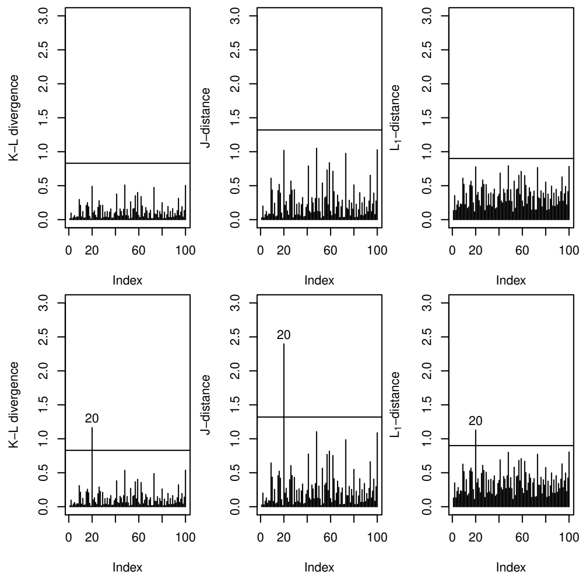

Figure 2.6: The q-divergence measures (K-L, J and L_1 distance) without perturbation (upper panel), and after perturbing subject ID #20 (lower panel) for the simulated data.

2.5 Conclusions

Motivated by the classical development of (Ospina and Ferrari, 2010), we developed a model for clustered responses in $[0, 1]$, and applied it to an interesting PrD dataset. Our model allows the parameters p_{0ij}, p_{1ij} and μ_{ij} to depend on covariates, leading to identifying covariates that are significant to explain disease-free, progressing with disease, and completely diseased tooth types. We also developed tools for outlier detection using q-divergence measures, and quantified their effect on the posterior estimates of the model parameters. Both simulation studies and real data application justify seeking an appropriate theoretical model over utilizing ad hoc data transformations for proportion data. Note that the proposition in Ospina and Ferrari (2010) (without any 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 and 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 model can be fitted using standard available software packages, such as `R` and `OpenBUGS`, with easy access to practitioners in the field.

It is of interest to investigate the presence of thick/heavy tails in the underlying ZOAB-RE proposition, and to model the random effect term b_i using robust alternatives (say, the t-density) over the normal density as in Figueroa-Zúniga et al. (2013). For our dataset, the results were very similar using a t-density, and hence we did not consider it any further.

Our current analysis considers clustered cross-sectional periodontal proportion data. Often, these study subjects can be randomized to dental treatments and subsequent longitudinal follow-ups, leading to a clustered-longitudinal framework, where one might be interested in estimating the profiles (both overall, and subject-level) in the proportion of diseased surfaces for the four tooth types with time. Our ZOAB-RE can certainly be extended to such situations with proper consideration to the GLMM REs specification. Other propositions available in the literature on modeling clustered (or longitudinal) proportion responses include simplex mixed-effects models (Qiu et al., 2008), robust transformation models (Song and 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 components of future research, and will be considered elsewhere.

Chapter 3

Augmented mixed models for clustered proportion data using the simplex distribution

Abstract
Proportional continuous data can be found in areas such as biological sciences, health, engineering, etc. This type of data, doubly bounded, assumes values in the interval $(0, 1)$ and for analysis, distributions such as logistic-normal, beta, beta-rectangular and simplex, among others, have been used. However, because in practical situations it is possible to observe, proportions, rates or percentages that are zero and/or one, these distributions cannot be used. To deal with this, we propose a regression model based on the simplex distribution that allows modelling the values zero and one simultaneously. For our analysis, we adopt a Bayesian framework and develop a Markov chain Monte Carlo algorithm to carry out the posterior analyses for longitudinal proportional data. Bayesian case deletion influence diagnostics based on the q-divergence measure and model selection criteria are also developed. We illustrated the proposed methodology through both simulation studies and real data to demonstrate the performance of our proposal.

Keywords Augmented distributions; Bayesian inference; MCMC; simplex distribution; q-divergence measures.

3.1 Introduction

Double-bounded data can be found in different areas such as biology, health sciences and engineering, among many others. Some of the strategies pointed out in the statistical literature to analyze this type of data are based on regression models combined with a particular data transformation such as the *logit* transformation. However, the use of nonlinear transformations may hinder the interpretation of the regression parameters. This situation can be overcome by considering probability distributions with double-bounded support, such as the beta, simplex (Barndorff-Nielsen and Jørgensen, 1991), and beta rectangular distributions (Hahn, 2008), which can be parameterized in terms of their mean. Other distributions, such as the logistic normal (Atchison and Shen, 1980) and the Kumaraswamy distribution

(Kumaraswamy, 1980) also have support in the unit interval. Nevertheless, the probability density function (*pdf*) of these distributions cannot be parameterized in terms of the means, limiting their use in regression analysis.

In this work, we focus on the simplex distribution to model proportions, rates or fractional data because its *pdf* presents a wide range of shapes including skewed, bimodal and multimodal ones. Based on this distribution, Song and Tan (2000) proposed a regression model relating the covariates with the mean via the logit link and assuming a fixed dispersion parameter. In that case, the the parameters are estimated through an extended version of the generalized estimating equations (GEE). Subsequently, Song et al. (2004) relaxed the condition over the dispersion parameter considering it to be heterogeneous as a function of the covariates through the logarithm link function. On the other hand, Qiu et al. (2008) derived the penalized quasi-likelihood (PQL) and restricted maximum likelihood (REML) (Breslow and Clayton, 1993), using the high-order multivariate Laplace approximation, which gives satisfactory maximum likelihood (ML) estimation of the model parameters in the simplex mixed-effects model. More recently, López (2013) presented a Bayesian approach for estimating the parameters in the simplex regression model where the response variable is confined in the interval $(0, 1)$.

In practical situations, proportions zero and/or one can be observed. Alternatives for the analysis of this type of data (in the interval $[0, 1]$) consider some transformations such as that proposed by Smithson and Verkuilen (2006). In this case, data values in the interval $[0, 1]$ are transformed to values in the interval $(0, 1)$. Once the transformation is applied, distributions like the beta or simplex, among others can be used. However, these transformations induce estimates with poor statistical properties even when small quantities of zeros and ones are observed (Galvis et al., 2014). For that reason and motivated by our data application, which includes zeros and ones, we propose a zero and one augmented simplex model (ZOAS), which allows us to deal with data in the interval $[0, 1]$ without the need for transformations.

After fitting the model, it is important to conduct sensitivity analysis to detect possible influential observations. An important approach to identify these influential cases is the case-deletion method introduced by Cook (1977). In the Bayesian context, Xie et al. (2014) investigated the Bayesian estimation and case influence diagnostics for the zero-inflated generalized Poisson regression model and Galvis et al. (2014) studied the Bayesian case-deletion diagnostics for the zero-and-one augmented beta random effects model. To the best of our knowledge, the Bayesian approach for drawing influence diagnostics in ZOAS random effects (ZOAS-RE) models has not been investigated in the literature. Therefore, an additional purpose of this work is to discuss and to develop some Bayesian influence diagnostic measures, based on the q-divergence, as proposed by Peng and Dey (1995), for the ZOAS-RE model. These Bayesian measures can be easily implemented with standard Bayesian software packages such as *OpenBUGS* (Thomas et al., 2006).

The paper is organized as follows. Section 3.2 presents some characteristics of the family of dispersion models and introduces the simplex distribution as an element of this family. In addition, the ZOAS regression model and the ZOAS-RE model are presented. Section 3.3 deals with the Bayesian inference, Bayesian model selection tools and case influence diagnostics for our proposed models. The application of the ZOAS-RE model is presented in Section 3.4 and simulation studies are presented in Section 3.5. Finally, some concluding remarks and avenues for further research are presented in Section 3.6.

3.2 Statistical model

3.2.1 Preliminaries

Dispersion models (DM) (Jørgensen, 1997) are a bi-parametric class of distributions whose elements have a *pdf* given by

$$f(y|\mu,\sigma^2) = a(y,\sigma^2)\exp\left\{-\frac{1}{2\sigma^2}d(y,\mu)\right\}, \tag{3.2.1}$$

where $E[Y]=\mu$, $\sigma^2>0$ is a dispersion parameter, $a(y,\sigma^2)>0$ does not depend on μ and the function $d(\cdot,\cdot)$ measures the discrepancy between the observed y and the expected μ. Through this function it is possible to identify each element belonging to the DM family of distributions. This function is called *unit deviance* if it satisfies $d(y,y)=0$ when $y=\mu$ and $d(y,\mu)\geq 0$ when $y\neq\mu$. Moreover, the *unit deviance* is said to be *regular* if it is twice continuously differentiable with respect to (y,μ), satisfying $\frac{\partial^2}{\partial\mu^2}d(y,\mu)=\frac{\partial^2}{\partial\mu^2}d(y,y)\Big|_{y=\mu}>0$. In this case (*regular unit deviance*), the variance function is defined as

$$V(\mu) = 2\left\{\frac{\partial^2}{\partial\mu^2}d(y,\mu)\Big|_{y=\mu}\right\}^{-1}.$$

Some well known distributions such as the normal, gamma, inverse normal, binomial and Poisson, among others, are particular cases of the DM models and, additionally, belong to a subclass of the DM family called exponential dispersion models (EDM), as previously proposed by Jørgensen (1987). Note that our work is focused on the simplex distribution introduced by Barndorff-Nielsen and Jørgensen (1991), which also belongs to the DM class of distributions. The simplex distribution is flexible and presents a wide variety of shapes, including some multimodal ones, which cannot be obtained by its counterpart, the beta distribution (see Figure 3.1). The *pdf* of a random variable Y following a simplex distribution, with mean μ and dispersion parameter σ^2, denoted by $S(\mu,\sigma^2)$, is given by

$$f(y|\mu,\sigma^2) = \{2\pi\sigma^2[y(1-y)]^3\}^{-1/2}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2 y(1-y)\mu^2(1-\mu)^2}\right\}, \tag{3.2.2}$$

where $0<y<1$, $0<\mu<1$, $\sigma^2>0$.

From the *pdf* (3.2.1), we have that $a(y,\sigma)=\{2\pi\sigma^2[y(1-y)]^3\}^{-1/2}$ and $d(y,\mu)=-\dfrac{(y-\mu)^2}{2\sigma^2 y(1-y)\mu^2(1-\mu)^2}$. It can be shown that $d(y,\mu)$ is a *regular unit deviance* and therefore the variance function for the simplex distribution is given by $V(\mu)=\mu^3(1-\mu)^3$.

Figure 3.1 shows several shapes of the simplex distribution for some values of μ and σ^2. Note that, when μ is close to zero (one) and the value of σ^2 is large, the mass of the simplex model is in the left (right) tail of the distribution. Moreover, when μ is close to 0.5 and the value of σ^2 is large, the *pdf* of the simplex distribution is bimodal, unlike the beta distribution, which has a unique mode.

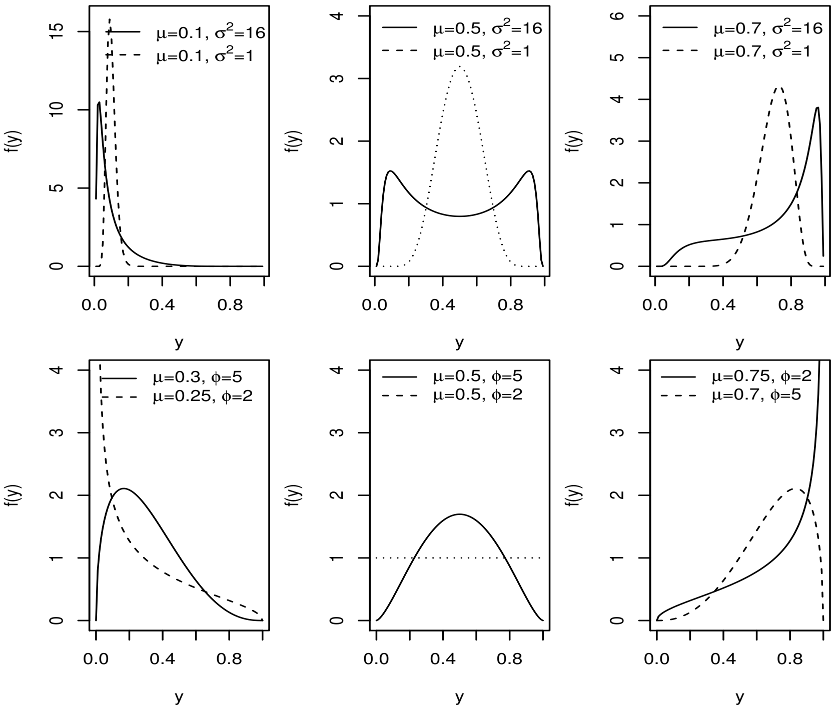

Figure 3.1: (Upper panel) *pdf* of the simplex distribution and (lower panel) *pdf* of the beta distribution with mean μ and precision parameter ϕ.

3.2.2 Simplex regression model

In order to define the simplex regression model, we consider the random variables $y_1, \ldots, y_n$ following the distribution given in (3.2.2). To relate the mean of the distribution with the covariate vector $\mathbf{x}_i$, $i = 1 \ldots, n$, we used a link function g_1 with domain in the interval (0,1) and range on the real line. This function is given by $g_1(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of regression parameters of dimension p where the first element of $\mathbf{x}_i$ is equal to one to accommodate the intercept. The dispersion parameter σ^2 can be considered invariant for all subjects as was adopted by Song and Tan (2000) and Qiu et al. (2008) or it can be regressed onto the covariates by using link functions such as the log, square root, etc, as was proposed by Song et al. (2004). The parameter estimation is conducted via ML in the classical context or using a Markov chain Monte Carlo (MCMC) scheme in the case of the Bayesian approach.

3.2.3 Zero-and-one augmented simplex random effects model

To deal with longitudinal data with observations in the $[0, 1]$ interval, the ZOAS-RE model is defined. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be n random vectors, where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ is a response vector of length n_i corresponding to the i-th subject. In order to define the ZOAS-RE model, it is assumed that conditional on the random effects $\mathbf{b}_i = (b_{i1}, \ldots, b_{iq})^\top$, the components Y_{ij} of $\mathbf{Y}_i$ are independent and distributed according to the ZOAS model whose *pdf* is given by

$$f(y_{ij}|p_{0ij},p_{1ij},\mu_{ij},\sigma^2)=\begin{cases} p_{0ij}, & \text{if } y_{ij}=0,\\ p_{1ij}, & \text{if } y_{ij}=1,\\ (1-p_{0ij}-p_{1ij})f(Y_{ij}=y_{ij}|\mu_{ij},\sigma^2), & \text{if } y_{ij}\in(0,1), \end{cases} \tag{3.2.3}$$

where $f(y_{ij}|\mu_{ij},\sigma^2)$ is as in (3.2.2), $j=1,\ldots,n_i$, $p_{0ij}>0$, $p_{1ij}>0$ and $p_{0ij}+p_{1ij}<1$. We denote by $\mathbf{Y}_{ij}|\mathbf{b}_i \sim ZOAS(\mu_{ij},\sigma^2,p_{0ij},p_{1ij})$ if the *pdf* of Y_{ij} is given as in (3.2.3). Note that, in this case, the model parameters can be regressed onto some covariates using appropriate link functions, as follows:

$$\begin{aligned} g_1(\boldsymbol{\mu}_i) &= \mathbf{X}_i^\top\boldsymbol{\beta}+\mathbf{Z}_i^\top\mathbf{b}_i,\\ g_2(\boldsymbol{p}_{\mathbf{0}i}) &= \boldsymbol{W}_{\mathbf{0}i}^{\top}\boldsymbol{\psi}\\ &\text{and}\\ g_3(\boldsymbol{p}_{\mathbf{1}i}) &= \boldsymbol{W}_{\mathbf{1}i}^{\top}\boldsymbol{\rho}, \end{aligned}$$

where $\boldsymbol{\mu}_i=(\mu_{i1},\ldots,\mu_{in_i})^\top$, $\mathbf{b}_i=(b_{i1},\ldots,b_{iq})^\top$ with $b_{i\cdot}\sim N(0,\sigma_b^2)$, $\boldsymbol{p}_{\mathbf{0}i}=(p_{0i1},\ldots,p_{0in_i})^\top$, $\boldsymbol{p}_{\mathbf{1}i}=(p_{1i1},\ldots,p_{1in_i})^\top$; $\mathbf{X}_i$, $\boldsymbol{W}_{\mathbf{0}i}$ and $\boldsymbol{W}_{\mathbf{1}i}$ are design matrices of dimension $p\times n_i$, $r\times n_i$ and $s\times n_i$ related to the fixed effects $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ and $\boldsymbol{\rho}$ respectively, and $\mathbf{Z}_i$ is the design matrix of dimension $q\times n_i$ related to the random effects vector $\mathbf{b}_i$. Link functions as the logit, probit or complementary log-log can be considered for g_1, g_2 and g_3. However, for the sake of interpretation, here we choose the logit function. As was mentioned previously, σ^2 (as well as the other parameters) can be regressed onto some covariates or considered invariant. In this work, we consider it invariant for all subjects. Finally, we define our ZOAS-RE model as $Y_{ij}|\mathbf{b}_i\sim ZOAS(\mu_{ij},\sigma^2,p_{0ij},p_{1ij})$, $i=1,\ldots,n$ and $j=1,\ldots,n_i$.

3.3 Bayesian Inference

3.3.1 Priors and posterior distributions

In order to complete the Bayesian specification, it is necessary to consider prior distributions for all the unknown model parameters. In this case, we use Normal multivariate distributions for the parameters $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ and $\boldsymbol{\rho}$. That is, $\boldsymbol{\beta}\sim \text{Normal}_p(\mathbf{0},\boldsymbol{\Sigma}_\beta^{-1})$, $\boldsymbol{\psi}\sim\text{Normal}_r(\mathbf{0},\boldsymbol{\Sigma}_\psi^{-1})$, $\boldsymbol{\rho}\sim\text{Normal}_s(\mathbf{0},\boldsymbol{\Sigma}_\rho^{-1})$. For the dispersion parameter, we considered a uniform distribution, which is $\sigma\sim\text{Unif}(0,a_1)$ with a large value for a_1. When the vector of probabilities $(p_0,p_1,1-p_0-p_1)^\top$ is considered invariant, we use the Dirichlet prior with hyperparameter $\boldsymbol{\alpha}^\top=(\alpha_1,\alpha_2,\alpha_3)$ and $\alpha_s\sim\text{Gamma}(1,0.001)$, $s=1,2,3$. The prior for the variance of RE is $\sigma_b\sim\text{Unif}(0,b_1)$, with a large positive value for b_1. Although multivariate specifications (multivariate zero mean vector with inverted-Wishart covariance) are certainly possible, we stick to simple (and independent) choices.

Posterior conclusions are based on the joint posterior distribution of all the model parameters (conditional on the data), and are obtained by combining the likelihood
$L(\boldsymbol{\Omega}|\mathbf{b},\mathbf{X},\mathbf{Z},\boldsymbol{W}_{\mathbf{0}},\boldsymbol{W}_{\mathbf{1}},\mathbf{y})$ given by

$$\prod_{i=1}^{n}\prod_{j=1}^{n_i}\left(p_{0ij}\right)^{I_{y_{ij}=0}}\left(p_{1ij}\right)^{I_{y_{ij}=1}}\left[(1-p_{0ij}-p_{1ij})f(y_{ij};\mu_{ij},\sigma^2)\right]^{I_{y_{ij}\in(0,1)}},$$

and the joint prior densities using the Bayes' rule. Thus, assuming a priori independence of the elements of the parameter vector, we can write

$$p(\boldsymbol{\Omega}, \mathbf{b}, \sigma_b | \mathbf{X}, \mathbf{Z}, \boldsymbol{W_0}, \boldsymbol{W_1}, \mathbf{y}) \propto L(\boldsymbol{\Omega} | \mathbf{b}, \mathbf{X}, \mathbf{Z}, \boldsymbol{W_0}, \boldsymbol{W_1}, \mathbf{y}) \times \pi(\boldsymbol{\Omega}, \mathbf{b}, \sigma_b),$$

where $\pi(\boldsymbol{\Omega}, \mathbf{b}, \sigma_b) = \pi_0(\boldsymbol{\beta})\pi_1(\boldsymbol{\psi})\pi_2(\boldsymbol{\rho})\pi_3(\sigma^2)\pi_4(\mathbf{b}|\sigma_b)\pi_5(\sigma_b)$ and $\pi_j(.), j = 0, \ldots, 5$ denotes the prior/hyperprior distributions for the model parameters as was described above.

The full conditional distributions necessary for the MCMC algorithm (a Metropolis-within-Gibbs algorithm) of the ZOAS-RE regression model are obtained as follows:

- $\pi(\boldsymbol{\beta} | \mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\beta})})$, is proportional to

$$\exp\left\{ -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \Sigma_\beta^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{\left[y_{ij} - (1 - y_{ij})A_{ij}\right]^2 (1 + A_{ij})^2}{2\sigma^2 y_{ij}(1 - y_{ij})A_{ij}^2} I_{\{y_{ij} \in (0,1)\}} \right\},$$

  where $A_{ij} = \exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta} + \mathbf{Z}_{ij}^\top \mathbf{b}_i)$.

- $\pi(\sigma^2 | \mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\sigma^2)})$ is a right truncated inverse gamma with truncation point a^2. That is $\sigma^2 | \mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\sigma^2)} \sim IGamma^+(a^2, N - 2, \sum_{i=1}^n \sum_{j=1}^{n_i} B_{ij})$, where $N = \sum_{i=1}^n n_i$ and $B_{ij} = \dfrac{(y_{ij} - \mu_{ij})^2}{2y_{ij}(1 - y_{ij})\mu_{ij}^2(1 - \mu_{ij})^2}$, $\mu_{ij} = \dfrac{A_{ij}}{1 + A_{ij}}$ and $y_{ij} \in (0, 1)$.

- $\pi(\mathbf{b}_i | \mathbf{y}, \sigma_b, \Omega)$ is proportional to

$$\exp\left\{ \frac{-1}{2\sigma_b^2} \sum_{k=1}^{q} b_{ik}^2 - \sum_{j=1}^{n_i} \frac{\left[y_{ij} - (1 - y_{ij})A_{ij}\right]^2 (1 + A_{ij})^2}{2\sigma^2 y_{ij}(1 - y_{ij})A_{ij}^2} I_{\{y_{ij} \in (0,1)\}} \right\} I_{\{b_{ik} \in \mathbb{R}\}}.$$

- When the probabilities p_0 and p_1 are regressed through some covariates, $\pi(\boldsymbol{\psi} | \mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\psi})})$ is proportional to

$$\exp\left\{ -\frac{1}{2}(\boldsymbol{\psi} - \boldsymbol{\psi}_0)^\top \Sigma_\psi^{-1} (\boldsymbol{\psi} - \boldsymbol{\psi}_0) \right\} \prod_{i=1}^{n} \prod_{j=1}^{n_i} C_{ij}^{I_{\{y_{ij}=0\}}} (1 - C_{ij} - D_{ij})^{I_{\{y_{ij} \in (0,1)\}}},$$

  where $C_{ij} = \text{logit}^{-1}(\boldsymbol{W_0}_{ij}^\top \boldsymbol{\psi})$ and $D_{ij} = \text{logit}^{-1}(\boldsymbol{W_1}_{ij}^\top \boldsymbol{\rho})$; and $\pi(\boldsymbol{\rho} | \mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\rho})})$, is proportional to

$$\exp\left\{ -\frac{1}{2}(\boldsymbol{\rho} - \boldsymbol{\rho}_0)^\top \Sigma_\rho^{-1} (\boldsymbol{\rho} - \boldsymbol{\rho}_0) \right\} \prod_{i=1}^{n} \prod_{j=1}^{n_i} D_{ij}^{I_{\{y_{ij}=1\}}} (1 - C_{ij} - D_{ij})^{I_{\{y_{ij} \in (0,1)\}}}.$$

- If p_0 and p_1 are considered invariant for all subjects, we have that $\pi(\mathbf{p} | \mathbf{y}, \mathbf{b}, \boldsymbol{\beta}, \sigma_b)$ is a Dirichlet distribution with parameters (ν_0, ν_1, ν_2), where $\nu_l = \alpha_l \sum_{i=1}^n \sum_{j=1}^{n_i} I_{y_{ij}=l}$, $l = 0, 1$ and $\nu_2 = \alpha_2 \sum_{i=1}^n \sum_{j=1}^{n_i} I_{y_{ij} \in (0,1)}$.

The relevant MCMC steps were implemented using the BRugs package (Ligges et al., 2009), which connects the R with the OpenBUGS software. After discarding 50000 burn-in samples, we used 50000 more samples (with a spacing of 50) from two 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 $\hat{\mathrm{R}}$, all of which are available in the R coda library (Cowles and Carlin, 1996). The associated BRugs code is available on request from the corresponding author.

3.3.2 Bayesian model selection and influence diagnostics

We use the conditional predictive ordinate (CPO) for our model selection derived from the posterior predictive distribution (*ppd*), and summarize these CPOs via the log pseudo-marginal likelihood (LPML) statistic (Carlin and Louis, 2008). Larger values of LPML indicate better fit. Owing to the instability of the harmonic-mean identity used for CPO computations (Raftery et al., 2007), we consider a more pragmatic route and compute the CPO (and 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. In addition, we also apply the expected AIC (EAIC), expected BIC (EBIC) (Carlin and Louis, 2008) and the DIC_3 (Celeux et al., 2006) criteria. The DIC_3 was used as an alternative to the usual DIC (Spiegelhalter et al., 2002) because of the ease of computation directly from the MCMC output, and also due to the mixture modeling framework. All these criteria abide by the 'lower is better' law, *i.e.*, the model producing the lowest value gets selected.

In addition, as a direct by product of the MCMC output, some influence diagnostic measures are developed to study the impact of outliers on mainly the fixed effects parameters due to data perturbation schemes based on case-deletion statistics (Cook and Weisberg, 1982), and the q-divergence measures (Csisz et al., 1967; Weiss, 1996) between posterior distributions. We consider three choices of these divergences, namely, the Kullback-Leibler (KL) divergence, the J-distance (symmetric version of the KL divergence), and the L_1-distance. We use the calibration method of Peng and Dey (1995) to obtain the cut-off values as 0.90, 0.83 and 1.32 for the L_1, KL and J-distances, respectively.

3.4 Application to periodontal disease proportions

We start this section with a brief description of the dataset. The Medical University of South Carolina (MUSC) performed a study in order to know the status and progression of clinical attachment level (CAL), a clinical marker of periodontal disease (*PrD*) among Gullah-speaking African-Americans with Type-2 diabetes. The main goal in that study is to identify covariates related with absence, limited presence and total presence of the disease. The dataset contain records on 28 teeth (considered full dentition, excluding the 4 third-molars) from 290 subjects, where the attention is focused on quantifying the extent and severity of *PrD* with respect to tooth-types. The observed 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', and therefore, a clustered data framework is generated, where each subject records four observations corresponding to the four tooth-types. Note that in this case, the response lies in the closed interval [0,1]; where 0 and 1 represent completely disease free and highly diseased cases, respectively. Missing teeth were considered 'missing due to *PrD*' where all sites for that tooth contributed to the diseased category. Subject-level covariates 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 includes current and past smokers. We also considered a tooth-level variable, representing each of the four tooth-types, with 'canine' as the baseline.

Due to the presence of a substantial number of 0's (114, 9.83%) and 1's (94, 8.10%), the use of a simplex regression might be inappropriate in this context. Consequently, we consider our proposed ZOAS-RE model in two different setups:

Model 1 $Y_{ij} \sim ZOAS-RE(\mu_{ij}, \sigma^2, p_{0ij}, p_{1ij})$,

Model 2 $Y_{ij} \sim ZOAS-RE(\mu_{ij}, \sigma^2, p_0, p_1)$,

where $\mu_{ij} = \text{logit}^{-1}(\mathbf{X}_{ij}^\top \boldsymbol{\beta} + b_i)$, $\mathbf{X}_{ij} = (1, \text{Gender}_i, \text{Age}_i, \text{HbA1c}_i, \text{Smoker}_i,$ $\text{Incisor}_{ij}, \text{Premolar}_{ij}, \text{Molar}_{ij})^\top$, β_0 is the intercept, $\beta_1, \ldots, \beta_7$ are the regression parameters, and b_i is the subject-level random effect term. For the parameters p_0 and p_1 we also use logit link functions, that is, $p_{0ij} = \text{logit}^{-1}(\boldsymbol{W_0}_{ij}^\top \boldsymbol{\psi})$ and $p_{1ij} = \text{logit}^{-1}(\boldsymbol{W_1}_{ij}^\top \boldsymbol{\psi})$ with $\boldsymbol{W_0}_{ij} = \boldsymbol{W_1}_{ij} = \mathbf{X}_{ij}$.

We also consider in the analysis the zero and one augmented beta model with random effects (ZOAB-RE) proposed by Galvis et al. (2014). This model uses the beta distribution parameterized as in Ferrari and Cribari-Neto (2004) to model data in $(0, 1)$. In this model, as in the ZOAS-RE model, the parameters p_0 and p_1 can be considered constants or regressed onto covariates. Therefore, as in the case of the ZOAS-RE model, we consider two natural competing models:

Model 3 $Y_{ij} \sim ZOAB-RE(\mu_{ij}, \phi, p_{0ij}, p_{1ij})$,

Model 4 $Y_{ij} \sim ZOAB-RE(\mu_{ij}, \phi, p_0, p_1)$.

In these models, we consider the same systematic part and link function used in models 1 and 2.

Although other link functions (such as probit, cloglog, etc) are available, here we restrict ourselves to the symmetric logit link. Those models can be compared using the model choice criteria described in Subsection 3.3.2. In the absence of historical data/experiments, our prior choices follow the specifications described in Subsection 3.3.1. In the case of the parameter ϕ in the ZOAB-RE model, we consider a Gamma prior, that is, $\phi \sim \text{Gamma}(0.1, 0.01)$.

Table 3.1 presents the DIC$_3$, LPML, EAIC and EBIC values calculated for models 1-4. Notice that, Model 1 (ZOAS-RE model with covariates in p_0 and p_1) outperforms the other models for all criteria. Therefore, we select Model 1 as our best model. Figure 3.2 plots the posterior parameter means and the 95% credible intervals (CIs) for regression onto μ (left panel), p_0 (middle panel) and p_1 (right panel) from the models 1 and 3. The gray intervals in this figure contain the zero value (the non-significant covariates), while the black intervals do not include the zero value (the significant ones at 5% level). In this figure, it can be noted that the significance of the covariates is similar in both models. However, this significance changes for the incisor covariate used to model μ, which is not significant under the Model 1.

Figure 3.2 (left panel) shows that for Model 1 the covariates Gender, Age and Premolar and Molar tooth type are related with the proportion of sites with PrD. Also, from Figure 3.2 (middle panel) it can be seen that the covariates Gender, Age and type of tooth are significant to explain the absence of PrD. Finally, the covariates Gender, Age and Molar are significant to explain disease completely (Figure 3.2, right panel). It is important to note the opposite form in which the covariates Gender, Age and Molar act on the parameters p_0 and p_1. That is, while the sign of the parameters related to Age, and tooth type are negatives to analyze absence of PrD (middle panel), they are positive to analyze complete presence of disease (right panel), indicating that older people has an odds less (greater) than younger people of be free (completely) of disease. Similar interpretation can be done for the covariate Molar.

Figure 3.3 presents the influence measures described in Subsection 3.3.2 for the ZOAS-RE model (upper panel) and for the ZOAB-RE model (lower panel). The ZOAS-RE model detected the subject with ID #174 as an influential observation, while the ZOAB-RE model detected four such subjects, #135, #159, #174 and #285. Hence, we conclude that the ZOAS-RE model is more robust than the ZOAB-RE model to accommodate outliers. In order to study the impact of subject #174 on the regression parameters, the ZOAS-RE model was adjusted by removing this subject from the dataset. The results did not show changes at the significance of the covariates. However, the coefficient associated with the molar covariate in the p_0 regression was strongly affected. This situation might be related to the absence of PrD for all tooth types of this subject.

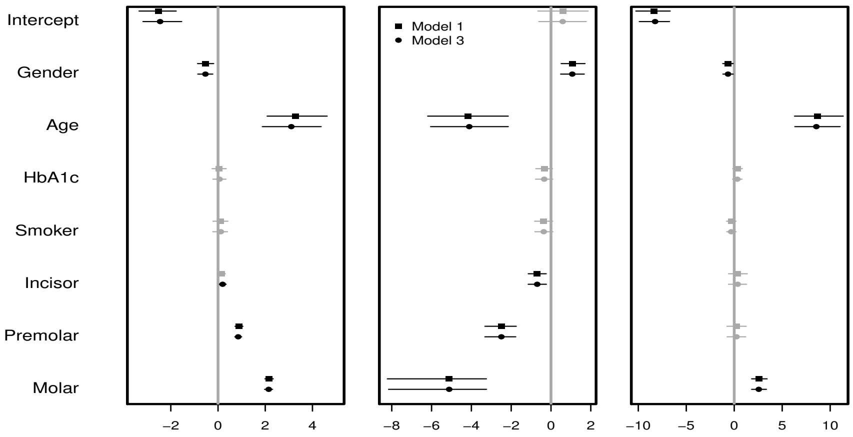

Figure 3.2: Posterior mean and 95% credible intervals (CI) of parameter estimates for mean not being zero or one (left panel), for p_0 (middle panel) and for p_1 (right panel) from Models 1 and 3. CIs that include zero are gray, those that do not include zero are black.

3.5 Simulation Studies

In this section, we conduct two simulation studies. In the first study, we analyze the performance of the ZOAS-RE and ZOAB-RE models and in the second one, we analyze the

	Model 1		Model 2		Model 3		Model 4	
DIC_3	**915.3**		1165.3		993.0		1243.5	
LPML	**-461.1**		-584.1		-500.5		-623.7	
EAIC	**917.8**		1154.9		992.7		1231.0	
EBIC	**1047.2**		1210.5		1124.2		1286.6	
Parameter	mean	SD	mean	SD	mean	SD	mean	SD
Intercept	-0.72^s	0.20	-0.71^s	0.21	-0.67^s	0.18	-0.67^s	0.18
Gender	-0.51^s	0.18	-0.51^s	0.17	-0.55^s	0.16	-0.54^s	0.17
Age	0.36^s	0.08	0.37^s	0.07	0.35^s	0.07	0.34^s	0.07
HbA1c	0.05	0.15	0.04	0.15	0.08	0.14	0.07	0.15
Smoker	0.11	0.16	0.10	0.17	0.11	0.16	0.12	0.15
Incisor	0.14	0.09	0.15	0.09	0.20^s	0.07	0.19^s	0.07
Premolar	0.89^s	0.09	0.89^s	0.09	0.85^s	0.07	0.85^s	0.07
Molar	2.17^s	0.09	2.17^s	0.09	2.15^s	0.08	2.14^s	0.08
σ^2	7.25	0.40	7.26	0.41	-	-	-	-
ϕ	-	-	-	-	7.60	0.43	7.63	0.42
p_0	-	-	0.098	0.009	-	-	0.098	0.009
p_1	-	-	0.081	0.008	-	-	0.081	0.008
σ_b^2	1.33	0.14	1.30	0.13	1.22	0.11	1.22	0.13

Table 3.1: Posterior parameter (mean) estimates and standard deviations (SD) obtained after fitting Models 1-4 to the periodontal data. ^{s} denotes a significant parameter.

effect of transforming the observed zero and one values on the Bayesian estimates of the regression parameters.

In both studies, the data was generated from a logistic normal distribution (Atchison and Shen, 1980), as follows. The location parameter μ_{ij} was generated as: $\mu_{ij} = \beta_0 + \beta_1 x_{ij} + b_i$, with, $b_i \sim N(0, \sigma_b^2)$, $i = 1, \ldots, n$, $j = 1, \ldots, 5$, indicating a cluster of size 5. Then, we generated a random variable T_{ij} following a normal distribution with mean μ_{ij} and variance 1. Next, we obtained the random variable Y_{ij} by applying the inverse logit of T_{ij}, that is, $Y_{ij} = \text{logit}^{-1}(T_{ij})$. This strategy generates values for y_{ij} in the interval $(0, 1)$. The final step is to allocate the 0's, 1's in the random sample $y_{ij} \in (0, 1)$. It is done by generating random samples from a multinomial distribution with probabilities vector $(p_{0ij}, p_{1ij}, 1 - p_{0ij} - p_{1ij})^\top$ in **simulation scheme 1** and $(p_0, p_1, 1 - p_0 - p_1)^\top$ in **simulation scheme 2**. The main goal in both studies is to compare the mean squared error (MSE), relative bias, and coverage probability for the regression parameter $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$.

Simulation scheme 1

In this case, the parameters p_{0ij} and p_{1ij} are modeled through covariates as $\text{logit}(p_{0ij}) = \psi_0 + \psi_1 W_{0ij}$ and $\text{logit}(p_{1ij}) = \rho_0 + \rho_1 W_{1ij}$, respectively. The explanatory variables x_{ij} and W_{1ij} are generated as an independent draws from a $\text{Unif}(0, 1)$ and $W_{0ij} = W_{0i}$ is generated as independent sample from a Bernoulli(0.8). The regression parameters and variance com-

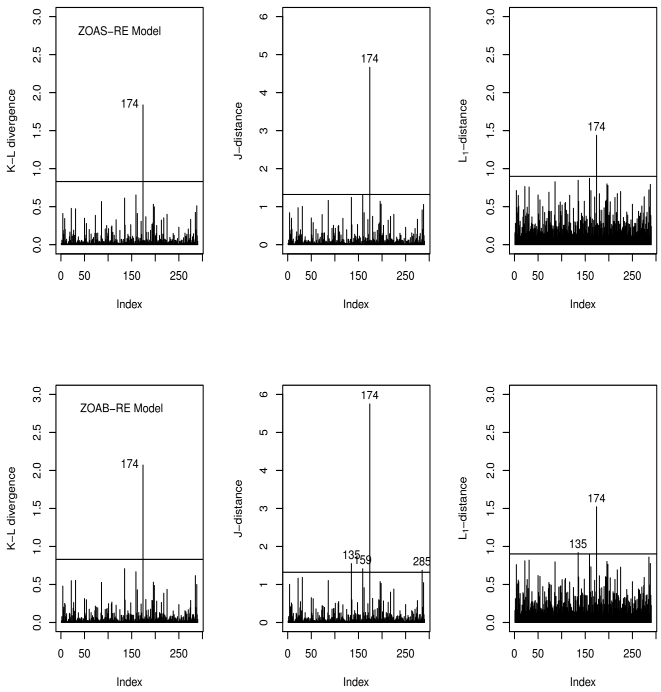

Figure 3.3: The q-divergence measures (K-L, J and L1 distances) for the application data using the ZOAS-RE model (upper panel) and the ZOAB-RE model (lower panel).

ponents are fixed at $\beta_0 = -0.5$, $\beta_1 = 0.5$, $\psi_0 = -1$, $\psi_1 = -1$, $\rho_0 = -1$, $\rho_1 = -1.5$ and $\sigma_b^2 = 2$, respectively.

We simulated 200 data sets with different sample sizes and fitted the ZOAS-RE and the ZOAB-RE model with both, p_0 and p_1 modeled as explained above. In all cases similar priors were chosen as those used in Section 3.4. Using the parameter vector $\boldsymbol{\theta} = (\beta_0, \beta_1, \psi_0, \psi_1, \rho_0, \rho_1, \sigma_b^2)$, with θ_s being an element of $\boldsymbol{\theta}$, we calculate the MSE as $\text{MSE}(\hat{\theta_s}) = \frac{1}{200}\sum_{i=1}^{200}(\hat{\theta}_{is} - \theta_s)^2$, the relative bias as Relative $\text{Bias}(\hat{\theta_s}) = \frac{1}{200}\sum_{i=1}^{200}\left(\frac{\hat{\theta}_{is}}{\theta_s} - 1\right)$, and the 95% coverage probability (CP) as $\text{CP}(\hat{\theta_s}) = \frac{1}{200}\sum_{i=1}^{200} I(\theta_s \in [\hat{\theta}_{s,LCL}, \hat{\theta}_{s,UCL}])$, 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.

Figure 3.4 presents the results obtained for the parameters β_0 and β_1. It can be observed that the ZOAS-RE model outperforms the ZOAB-RE model when the relative bias and CP are analyzed. However, both models exhibit similar MSE. The results related to the parameters ψ_0, ψ_1, ρ_0 and ρ_1 shown equal performance for both models, although this does not occur with the variance of the random effect, where the ZOAS-RE model outperforms the ZOAB-RE model. In the ZOAS-RE model, the absolute relative bias is close to 20% for all sample sizes, while in the ZOAB-RE model the relative bias is close to 40%. Also, the MSE for the ZOAS-RE model decreases when the sample size increases. This is 0.31 for $n = 50$ and 0.21 for $n = 200$. In the case of the ZOAB-RE model, the MSE remains around 0.60. Finally, the CP is greater for the ZOAS-RE model.

Further we fit the ZOAS-RE and ZOAB-RE models without covariates on p_0 and p_1. It is possible to note that the performance of the relative bias, MSE and CP of the parameters β_0, β_1 and σ_b^2 of these models are very close to those obtained considering covariates on p_0 and p_1. Thus, we can conclude that considering p_0 and p_1 as constants, the Bayesian estimates of the regression coefficients β_0 and β_1 are not affected.

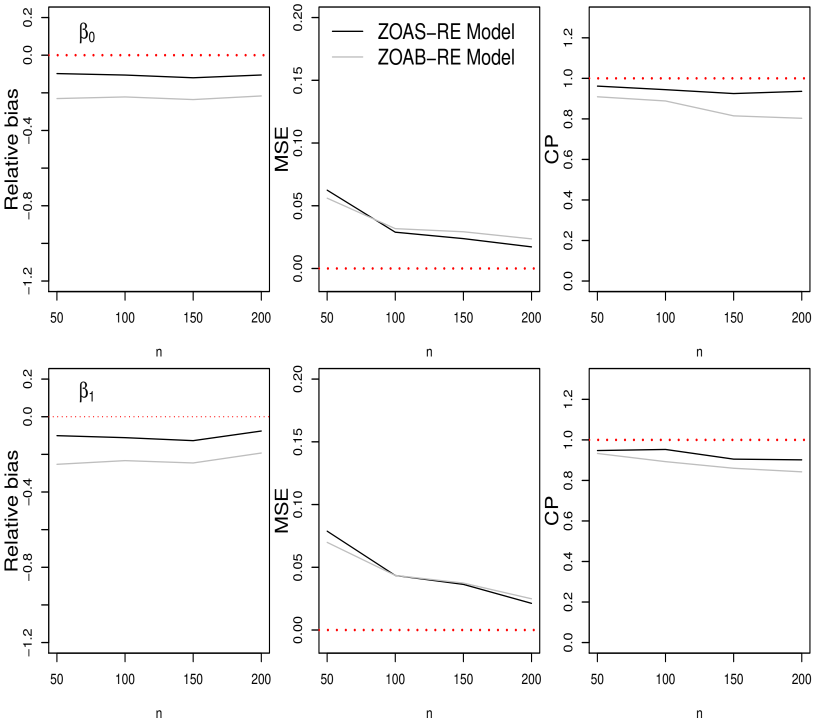

Figure 3.4: Relative bias, MSE and CP of β_0 and β_1 after fitting the ZOAS-RE (black line) and ZOAB-RE (gray line) models for the simulated data.

Simulation scheme 2

In this study, the parameters p_0 and p_1 are considered constants across all subjects by assuming the probability values $p_0 = 1\%, p_1 = 1\%$ in **case a** and $p_0 = 10\%, p_1 = 8\%$ in **case b**. Furthermore, the regression parameters and variance components are fixed at $\beta_0 = -0.5$, $\beta_1 = 0.5$ and $\sigma_2^b = 0.8$. Using these values, we generated 200 datasets following the scheme above and we fit our ZOAS-RE model and the simplex regression model (S-RE model) after transformation. The transformation used was that proposed by Smithson and Verkuilen (2006), where 0's and 1's are approximated by $1/2N$ and $(2N-1)/2N$, respectively, and N is the total number of observations. The main goal of this study is to analyze the effect of this transformation on the estimates of the regression parameters $\boldsymbol{\beta}$.

Figure 3.5 displays the results of Relative Bias, MSE and CP for the estimates of β_0 and β_1 in **case a** (upper panel) and **case b** (lower panel). As can be seen from this figure, even though the transformation of 0's and 1's is not prominent, there is a strong impact in the statistical properties of the regression estimates. The results obtained in the variances of RE are described next (figure not shown). In the ZOAS-RE model, the Relative Bias of σ_2^b is around of 20% for all sample sizes, while in the simplex counterpart, it starts at 0.34 for

$n = 50$ and greaches 1.34 for $n = 200$, indicating an increasing of the Relative Bias when the sample size increases. The MSE for this parameter is close to 5% in the ZOAS-RE model, but in the simplex model (transformed) it increases from 0.15 for $n = 50$ to 0.85 for $n = 200$. Also, when analyzing the CP, the ZOAS-RE model outperforms the (transformed) simplex model.

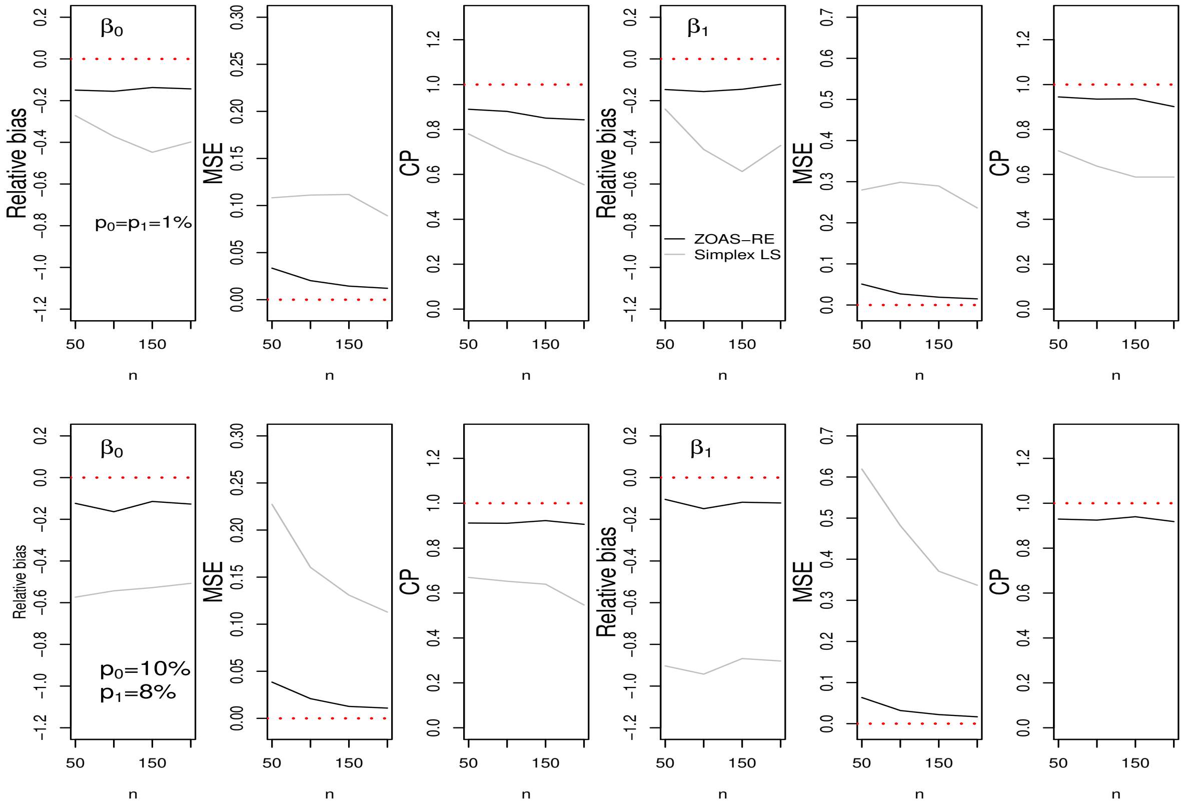

Figure 3.5: Relative Bias, MSE and CP of β_0 and β_1 after fitting the ZOAS-RE (black line) and S-RE (gray line) models using $(p_0, p_1)^\top = (1\%, 1\%)$ (upper panel) and $(10\%, 8\%)$ (lower panel).

3.6 Conclusions

This article proposes a Bayesian random effect model based on the simplex distribution for modeling data in the interval $[0, 1]$. The versatility of this class to model correlated data in the interval $[0, 1]$ has not been explored elsewhere, and this is our major contribution. Simulation studies reveal good consistency properties of the Bayesian estimates when compared with the beta regression counterpart, as well as, high performance of the model selection techniques to pick the appropriately fitted model. We also apply our method to a data set from periodontal disease conducted at the Medical University of South Carolina (MUSC) to illustrate how the procedures can be used to evaluate model assumptions, identify outliers and obtain unbiased parameter estimates. Although our modeling is primarily motivated from periodontal disease data, it can be easily applied to other datasets, since the models considered in this article have been fitted using standard available software packages, like R and OpenBUGS. This makes our approach easily accessible to practitioners of many fields of research. This paper complements the recently published work of Galvis et al. (2014), which also considers Bayesian estimation and inference of this kind of data by using the beta distribution (Ferrari and Cribari-Neto, 2004).

The models developed here do not consider skewness in the random effects and their robustness can be seriously affected by the presence of skewness and heavy tails in the random effects. Recently, Lachos et al. (2009) adopted a Markov chain Monte Carlo approach to draw Bayesian inferences in linear mixed models with multivariate skew-normal (SNI) distributions in the random effects. Therefore, it would be a worthwhile task to investigate the applicability of a Bayesian treatment in the context of ZOAS-RE models with SNI distributions. Incorporating measurement error in covariates (Carrasco et al., 2014) within our robust framework is also part of our future research.

Chapter 4

Augmented mixed models for clustered proportion data

Abstract
Often in biomedical research, we deal with continuous (clustered) proportion responses ranging between zero and one quantifying the disease status of the cluster units. Interestingly, the study population might also consist of relatively disease-free as well as highly diseased subjects, contributing to proportion values in the interval [0, 1]. Regression on a variety of parametric densities with support lying in $(0, 1)$, such as beta regression, can assess important covariate effects. However, they are deemed inappropriate due the presence of zeros and/or ones. To evade this, we introduce a class of general proportion density (GPD), and further augment the probabilities of zero and one to this GPD, controlling for the clustering. Our approach is Bayesian, and presents a computationally convenient framework amenable to available freeware. Bayesian case-deletion influence diagnostics based on q-divergence measures are automatic from the MCMC output. The methodology is illustrated using both simulation studies and application to a real dataset from a clinical periodontology study.
Keywords: Augment; Bayesian; Dispersion models; Kullback-Leibler divergence; Proportion data; Periodontal disease.

4.1 Introduction

Continuous proportion data (expressed as percentages, proportions, and rates), such as the percent decrease in glomerular filtration rate at various follow-up times since baseline Song and Tan (2000); Kieschnick and McCullough (2003) are routinely analyzed in medicine and public health. Because the responses are confined in the open interval $(0, 1)$, one might be tempted to use the logistic-normal model (Aitchison, 1986) with Gaussian assumptions for logit-transformed proportion responses. However, covariate effects interpretation are not straightforward because the logit link is no longer preserved for the expected value of the response. Alternatively, to tackle this, the beta (Cepeda-Cuervo, 2001; Ferrari and Cribari-Neto, 2004), beta rectangular (BRe) (Hahn, 2008) and simplex (Barndorff-Nielsen and Jørgensen, 1991) distributions (all with common support within the open unit interval), and their corresponding regressions were proposed under a generalized linear model (GLM) framework.

The flexible beta density (Johnson et al., 1994) can represent a variety of shapes, accounting for uncorrectable non-normality and skewness (Smithson and Verkuilen, 2006) in the context of bounded proportion data. The beta regression (BR) reparameterizes the associated beta parameters, connecting the response to the data covariates through suitable link functions (Ferrari and Cribari-Neto, 2004). Yet, the beta density does not accommodate tail-area events, or flexibility in variance specifications (Bayes et al., 2012). To accommodate this, the BRe density Hahn (2008), and associated regression modelsBayes et al. (2012) were considered under a Bayesian framework. Note, the BRe regression includes the (constant dispersion) BRFerrari and Cribari-Neto (2004), and the variable dispersion BR Smithson and Verkuilen (2006) as special cases. The simplex regressionSong and Tan (2000) is based on the simplex distribution from the dispersion family (Jørgensen, 1997), assumes constant dispersion, and uses extended generalized estimating equations for inference connecting the mean to the covariates via the logit link. Subsequently, frameworks with heterogenous dispersion (Song et al., 2004), and for mixed-effects models (Qiu et al., 2008) were explored. Yet, their potential were limited to proportion responses with support in $(0, 1)$.

A clinical study on periodontal disease (PrD) conducted at the Medical University of South Carolina (MUSC)(Fernandes et al., 2006) motivates our work. The clinical attachment level (CAL), a clinical marker of PrD was measured at each of the 6 sites of a subject's tooth, and we were interested to assess covariate-response relationships on 'tooth-type specific (such as incisors, canines, pre-molars and molars) proportion of diseased sites' to determine the status of PrD. Figure 4.1 (left panel) plots the raw (unadjusted) density histogram of the proportion responses, packed over all subjects and tooth-types. The responses are in the closed interval $[0, 1]$ where 0 and 1 represent 'completely disease free', and 'highly diseased' cases, respectively. For a simple parametric treatment to this data, one might be tempted to use one of the three distributions mentioned above after possible transformation Smithson and Verkuilen (2006) of the response from $[0, 1]$ to the interval $(0, 1)$. These ad hoc rescalings might work out for small proportions of 0's and 1's, but the sensitivity on parameter estimates can be considerable as the proportions increase. Transformations, in general, are not universal. In addition, presence of clustering (tooth-sites within mouth) brings in an extra level of heterogeneity, and these transformations which are usually applied component-wise may not guarantee a tractable (multivariate) joint distribution Jara et al. (2008). At this stage, we desire an appropriate theoretical model capable of handling all these challenges, yet avoiding data transformations.

Note that the beta, BRe and simplex densities (and their regressions) present a noticeable analytic difference in their probability density function (pdf) specification. Motivated by these differences and the flexibility they provide, we seek to combine them into a new (parametric) class of density called the general proportion density (GPD), where these three popular models appear as particular cases. In this context, our paper generalizes the recent augmented beta proposition Galvis et al. (2014). Next, we extend this GPD to a regression setup for independent responses in $(0, 1)$. Finally, for a unified (regression) framework for clustered responses in $[0, 1]$, we propose a generalized linear mixed model (GLMM) framework by augmenting the probabilities of occurrence of zeros, ones or both to the standard GPD regression model via an augmented GPD random effects (AugGPD-RE) model. Our inferential framework is Bayesian, and can be easily handled using freeware like `OpenBUGS`.

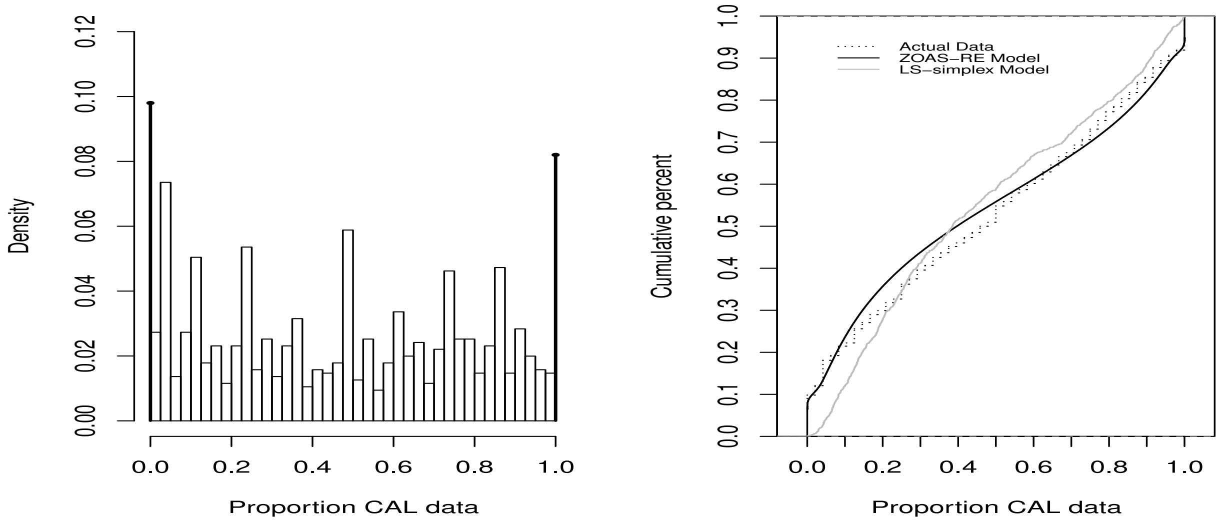

Figure 4.1: Periodontal proportion data. The (raw) density histogram combining subjects and tooth-types are presented in the left panel. The empirical cumulative distribution function of the real data, and that obtained after fitting the ZOAS-RE and the LS-simplex models appear in the right panel.

Furthermore, case-deletion and local influence diagnostics (Peng and Dey, 1995) to assess outlier effects are immediate from the Markov chain Monte Carlo (MCMC) output.

The rest of the article is organized as follows. Section 2 formulates the GPD and the augmented GPD class of density as well as some useful statistical properties. Section 3 develops the Bayesian estimation framework for the AugGPD-RE regression model and related diagnostics. Application to the motivating PD data appear in Section 4. Section 5 presents simulation studies to compare finite-sample performance of parameter estimates among the GPD class members, and also under model misspecification. Finally, some concluding statements appear in Section 6.

4.2 General proportion density

We start with the definition of proportion density (PD) models, and then proceed to establish the GPD density class.

Definition 1. *A random variable (rv) ξ with support in the unit interval $(0, 1)$ belongs to the class of PD with parameters λ and ϕ if it can be expressed as*

$$g_1(\xi; \lambda, \phi) = a_1(\lambda, \phi)a_2(\xi, \phi) \exp\{-\phi a_3(\xi, \lambda)\}, \quad \phi > 0, \; \lambda \in (0, 1), \qquad (4.2.1)$$

where $E[\xi] = \lambda$, and $a_s(\cdot, \cdot)$, $s = 1, 2, 3$ are real-valued functions with $a_1, a_2 \geq 0$, and a_3 taking value on the real line. We use the notation $\xi \sim PD(\lambda, \phi)$ to represent ξ a member of the PD class defined in (4.2.1). Following Jørgensen (1997), if $a_3(\xi, \lambda)$ in (4.2.1) is continuous and *twice differentiable function with respect to ξ and λ and is non-zero, the variance function is* $V(\lambda) = -\left(\dfrac{\partial^2 a_3(\xi, \lambda)}{\partial\lambda\partial\xi}\right)^{-1}\Bigg|_{\xi=\lambda}$.

Next, consider the density of the rv X following the 2-component mixture $X = \eta U + (1-\eta)\xi$, where $\eta \in [0, 1]$ is a mixture parameter and U a Uniform$(0, 1)$ rv distributed independently of ξ with pdf in (4.2.1). Then, X follows the general proportion density (GPD), i.e., $X \sim$ GPD(η, λ, ϕ) with the pdf given by

$$g(X; \eta, \lambda, \phi) = \eta + (1 - \eta)g_1(X; \lambda, \phi), \tag{4.2.2}$$

where g_1 is as defined in (4.2.1). Note that for $\eta = 1$, the GPD reduces to the uniform distribution, and for $\eta = 0$ we retrieve the PD class of distributions. The mean and variance of X are $\mu = E[X] = \eta/2 + (1 - \eta)E[\xi]$, $\sigma^2 = \text{Var}(X) = \frac{\eta}{12} + (1 - \eta)^2\text{Var}(\xi)$, respectively.

4.2.1 Densities in the GPD class

The GPD class includes the beta, simplex, and the BRe densities with support in the interval (0,1), and can be used to model proportion data. These are described in the propositions below with their respective pdf's presented in Appendix A.

Proposition 1. *The beta density (Ferrari and Cribari-Neto, 2004) reparametrized in terms of μ (the mean) and of ϕ (the precision parameter) belongs to the GPD class of distributions with its variance function given by $V(\mu) = \mu(1 - \mu)$.*

Proof. *In (4.2.2), consider $\eta = 0$, $\lambda = \mu$ and* $g_1(x; \mu, \phi) = \dfrac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}x^{\mu\phi-1}(1 - x)^{(1-\mu)\phi-1}$*, such that* $a_1(\mu, \phi) = \dfrac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}$, $a_2(x, \phi) = \left(x(1-x)^{1-\phi}\right)^{-1}$ *and* $a_3(x, \mu) = \mu \log \frac{1-x}{x}$. *Then, the variance functionFerrari and Cribari-Neto (2004) (from Definition 1) is*

$$V(\mu) = -\left(\frac{\partial^2 a_3(x, \mu)}{\partial\mu\partial x}\right)^{-1}\Bigg|_{x=\mu} = -\left(\frac{-1}{x(1-x)}\right)^{-1}\Bigg|_{x=\mu} = \mu(1 - \mu)$$

.

Proposition 2. *The simplex distribution (Barndorff-Nielsen and Jørgensen, 1991) with parameters μ and ϕ belongs to the GPD class with the variance function given by $V(\mu) = \mu^3(1 - \mu)^3$.*

Proof. *In (4.2.2), consider $\eta = 0$, $\lambda = \mu$ and* $g_1(x; \mu, \phi) = \dfrac{\sqrt{\phi}}{\sqrt{2\pi}\,(x(1-x))^{3/2}}$ $\times \exp\left\{-\phi\dfrac{(x-\mu)^2}{2x(1-x)\mu^2(1-\mu)^2}\right\}$*, such that* $a_1(\mu, \phi) = 1$, $a_2(x, \phi) = \dfrac{\phi}{\sqrt{2\pi}\,(x(1-x))^{3/2}}$ *and* $a_3(x, \mu) = \dfrac{(x-\mu)^2}{2x(1-x)\mu^2(1-\mu)^2}$. *Then, the variance functionJørgensen (1997) is given by*

$$V(\mu) = -\left(\tfrac{\partial^2 a_3(x,\mu)}{\partial\mu\partial x}\right)^{-1}\Bigg|_{x=\mu} = -\left(\tfrac{-1}{x^3(1-x)^3}\right)^{-1}\Bigg|_{x=\mu} = \mu^3(1 - \mu)^3.$$

Proposition 3. *The BRe density (Hahn, 2008) with parameters η, λ and ϕ belongs to the GPD class of distributions.*

Proof. *The proof follows from (4.2.2), considering $\eta > 0$ and $g_1(x; \lambda, \phi)$ as in Proposition 1, replacing μ by λ. However, the BRe density is a mixture of a uniform and a beta density (see Appendix A in the supplementary material) and a closed form expression of the variance function is not available.*

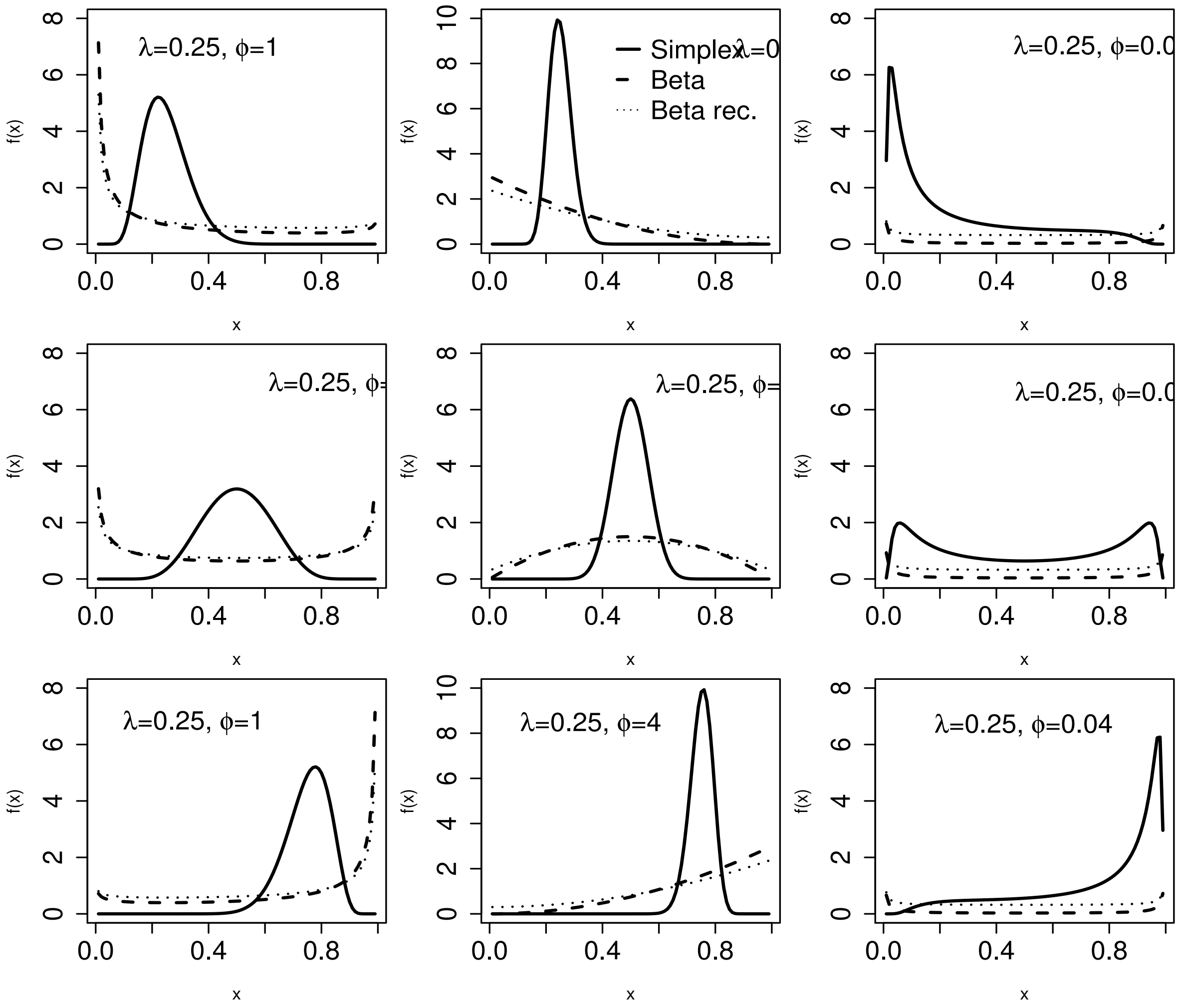

Figure 4.2: Plots of the simplex, beta and the beta rectangular densities for various choices of λ and ϕ. For the beta rectangular density, we choose $\eta = 0.3$.

For a more appealing pictorial comparison, Figure 4.2 plots the simplex, beta and the BRe densities for various choices of λ and ϕ. Note that λ close to zero (one) leads to a large mass in the left (right) tails for all cases. The simplex density is relatively smooth for $\phi = 1$, and becomes more spiked for $\phi = 4$. The beta and the BRe shapes are very similar for all panels when η is moderate ($= 0.3$), as in our case. However, one observes tail behavior for the BRe compared to the beta when η gets closer to 1 (plots not shown here). From the plots, it is clear that the simplex density is more flexible than the two competitors. It is capable of capturing various shapes of the underlying proportion data density in $(0, 1)$, even in situations (say, small ϕ) where the popular beta density may be far from the ground truth. However, a major shortcoming of these densities is that they are not appropriate for modeling datasets containing proportion responses at the extremes (i.e., 0, or 1, or both). We seek to address this via an augmented GPD framework defined as follows:

Definition 2. *The pdf of a rv Y with support in the interval $[0, 1]$ belongs to the augmented GPD class if it has the form*

$$f(y; \eta, \lambda, \phi, p_0, p_1) = p_0 I_{\{y=0\}} + p_1 I_{\{y=1\}} + (1 - p_0 - p_1) g(y; \eta, \lambda, \phi) I_{\{y\in(0,1)\}}, \tag{4.2.3}$$

where $I_{\{A\}}$ is the indicator function of the set A; $g(\cdot)$ is as defined in Equation (4.2.2) and p_0, $p_1 \geq 0$, with $p_0 + p_1 < 1$.

From (4.2.3), the expectation and variance of Y are, respectively, $E[Y] = p_1 + (1 - p_0 - p_1)\mu = \delta$ and $\text{Var}(Y) = p_1(1-p_1)+(1-p_0-p_1)[\sigma^2-2p_1\mu+(p_0+p_1)\mu^2]$, where μ and σ^2 are as in Definition 1. Note, the augmented GPD class defined in (2) reduces to the GPD class when p_0 and p_1 are simultaneously equals to zero. When $p_0 > 0$ and $p_1 = 0$ we have the zero augmented GPD class, and for $p_0 = 0$ and $p_1 > 0$ we have the one augmented GPD class. Finally, when $p_0 > 0$ and $p_1 > 0$, we have the more general zero-one augmented GPD class. Motivated by the PrD data, we are particularly interested in the following three subfamilies of the augmented GPD class, corresponding to the densities specified in Subsection 2.1

- Zero-one augmented beta (ZOAB) density, if $\eta = 0$ and $g_1(\cdot)$ the beta density
- Zero-one augmented simplex (ZOAS) density, if $\eta = 0$ and $g_1(\cdot)$ the simplex density
- Zero-one augmented beta rectangular (ZOABRe) density, if $\eta > 0$ and $g_1(\cdot)$ the beta density

4.3 Model development and Bayesian inference

4.3.1 GPD regression model

Let $Y_1, \ldots, Y_n$ be n independent rv's such that $Y_i \sim \text{GPD}(\eta_i, \lambda_i, \phi_i)$. **C**onsider that $\mu_i = \eta_i/2+(1-\eta_i)\lambda$ is directly modeled through covariates as $g_1(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$ where g_1 is a adequate link function with counterdomain the real line, $\boldsymbol{\beta}$ is the vector of regression parameters with the first element of $\mathbf{x}_i$ being 1. However, μ_i is a function of the mixture parameter η_i and λ, which leads to a restricted parametric space of η_i, defined as $0 < \eta_i < |2\mu_i - 1|$ that is dependent on μ_i. Hence, for a more appropriate regression framework that connects Y to covariates, we work with the reparameterization proposed in Bayes et al. (2012), and define $\alpha_i \in [0, 1]$ such that $\alpha_i = \dfrac{\eta_i}{1 - (1 - \eta_i)|2\lambda_i - 1|}$. Henceforth, the GPD class is parameterized in terms of μ_i, α_i and ϕ_i.

The parameters ϕ_i and α_i can be assumed constants, or regressed onto covariates through convenient link functions. For μ_i and α_i, link functions such as, logit, probit or complementary log-log can be used. Finally, for ϕ_i, the log, square-root, or identity link functions can be considered. Parameter estimation can follow either the (classical) maximum likelihood (ML), or the Bayesian route through MCMC methods.

4.3.2 Augmented GPD random effects model

The augmented GPD model described in (4.2.3) is only appropriate for independent responses in $(0, 1)$. To accommodate clustering (as in our case) or longitudinal subject-specific profiles, we proceed with the augmented GPD random effects (henceforth, AugGPD-RE) model. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be n independent continuous random vectors, where $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ is the vector of length n_i for the sample unit i, with the components $y_{ij} \in \zeta$, where ζ is an element of the set $\{[0, 1), (0, 1], [0, 1]\}$. Thus, under the AugGPD-RE model, the parameters μ_{ij}, p_{0ij} and p_{1ij} can be connected with covariates through suitable link functions as

$$\begin{aligned} g_1(E[\mathbf{Y}_i|\mathbf{b}_i]) = g_1(\boldsymbol{\mu}_i) = \mathbf{X}_i^\top \boldsymbol{\beta} + \mathbf{Z}_i^\top \mathbf{b}_i, \qquad (4.3.1) \\ g_2(\boldsymbol{p_0}_i) = \boldsymbol{W_0}_i^\top \boldsymbol{\psi}, \qquad (4.3.2) \\ g_3(\boldsymbol{p_1}_i) = \boldsymbol{W_1}_i^\top \boldsymbol{\rho}, \qquad (4.3.3) \end{aligned}$$

where $\mathbf{X}_{ij}$, $\boldsymbol{W_0}_{ij}$ and $\boldsymbol{W_1}_{ij}$ correspond to the j-th column from the design matrices $\mathbf{X}_i$, $\boldsymbol{W_0}_i$ and $\boldsymbol{W_1}_i$ of dimension $p \times n_i$, $r \times n_i$ and $s \times n_i$, related with the i-th unit sample, corresponding to the vectors of fixed effects $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_r)^\top$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_s)^\top$, respectively, 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, g_2 and g_3 remain the same as for μ_i and α_i in Subsection 4.3.1. For purpose of interpretation, we focus on the logit link. In this work, we consider ϕ and α as constants despite those parameters can also be regressed onto covariates through suitable link functions. Also, to avoid over-parameterization, the probabilities p_{0ij} and p_{1ij} are free of REs, however, both could be considered constants across subjects. Finally, we denote our AugGPD-RE model as $Y_{ij} \sim \text{AugGPD-RE}(p_{0ij}, p_{1ij}, \mu_{ij}, \alpha, \phi)$ $i = 1, \ldots, n,$ $j = 1, \ldots, n_i$.

Let $\boldsymbol{\mathcal{D}} = (\mathbf{X}_i, \boldsymbol{W_0}_i, \boldsymbol{W_1}_i, \mathbf{Z}_i, \mathbf{y})^\top$ be the full observed data and $\boldsymbol{\Omega} = (\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\rho}, \phi, \alpha)^\top$ be the parameter vector in the AugGPD-RE model. The joint data likelihood, conditional on the random-effects $\mathbf{b}_i$, $L(\boldsymbol{\Omega}; \boldsymbol{\mathcal{D}}, \mathbf{b})$ is given by

$$L(\boldsymbol{\Omega}; \mathbf{b}, \boldsymbol{\mathcal{D}}) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} {p_{0ij}}^{I_{y_{ij}=0}} {p_{1ij}}^{I_{y_{ij}=1}} \Big[(1 - p_{0ij} - p_{0ij}) g(y_{ij}; \alpha, \mu_{ij}, \phi)\Big]^{I_{y_{ij} \in (0,1)}}, \qquad (4.3.4)$$

where $p_{0ij} = \text{logit}^{-1}(\boldsymbol{W_0}_{ij}^\top \boldsymbol{\psi})$, $p_{1ij} = \text{logit}^{-1}(\boldsymbol{W_1}_{ij}^\top \boldsymbol{\rho})$, I is an indicator function, and g is given by

$$g(y_{ij}; \alpha, \mu_{ij}, \phi) = \eta_{ij} + (1 - \eta_{ij}) a_1(\lambda_{ij}, \phi) a_2(y_{ij}, \phi) \exp\left\{-\phi a_3(y_{ij}, \lambda_{ij})\right\}, \qquad (4.3.5)$$

with $\eta_{ij} = \alpha(1 - 2|\mu_{ij} - \frac{1}{2}|)$, $\lambda_{ij} = \dfrac{\mu_{ij} - \frac{\eta_{ij}}{2}}{1 - \eta_{ij}}$ and $\mu_{ij} = \text{logit}^{-1}\left(\mathbf{X}_{\mu_{ij}}^\top \boldsymbol{\beta} + \mathbf{Z}_{b_{ij}}^\top \mathbf{b}_i\right)$.

Although ML estimation of $\boldsymbol{\Omega}$ is certainly feasible using standard softwares such as (e.g., `SAS`, `R`, etc), we seek a Bayesian treatment here. The Bayesian approach accommodates full parameter uncertainty through appropriate choice of priors choices, 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.

4.3.3 Priors and posterior distributions

In order to complete the Bayesian specification, we need to consider prior distributions for all the unknown model parameters. In particular, we specify practical weakly informative prior opinion on the fixed effects regression parameters $\boldsymbol{\beta}$, $\boldsymbol{\psi}$, $\boldsymbol{\rho}$, ϕ (dispersion parameter), α, and the random effects $\mathbf{b}_i$. In general, for the regression components, we can assume $\boldsymbol{\beta} \sim \text{Normal}_p(\mathbf{0}, \boldsymbol{\Sigma}_\beta^{-1})$, $\boldsymbol{\psi} \sim \text{Normal}_r(\mathbf{0}, \boldsymbol{\Sigma}_\psi^{-1})$, $\boldsymbol{\rho} \sim \text{Normal}_s(\mathbf{0}, \boldsymbol{\Sigma}_\rho^{-1})$. A Uniform$(0,1)$ densityBayes et al. (2012) was adopted as prior for α. Prior on each element of $\mathbf{b}_i$ are $N(0, \sigma_b^2)$, where $\sigma_b \sim \text{Uniform}(0, c_1)$, the usual GelmanGelman (2006) specification. The prior on ϕ for the specific models in Subsection 2.1 were chosen as follows:

(i) *Beta and BRe models*: $\phi \sim \text{Gamma}(a, c)$, with small positive values of a and c ($c \ll a$).

(ii) *Simplex model*: $\phi^{-1/2} \sim \text{Uniform}(0, a_1)$, with large positive value for a_1.

Assuming the elements of the parameter vector to be independent, the posterior conclusions are obtained combining the likelihood in (4.3.4), and the joint prior densities, given by

$$p(\boldsymbol{\Omega}, \mathbf{b}, \sigma_b | \boldsymbol{\mathcal{D}}) \propto L(\boldsymbol{\Omega}; \boldsymbol{\mathcal{D}}) \times \pi(\boldsymbol{\Omega}, \mathbf{b}, \sigma_b),$$

where $\pi(\boldsymbol{\Omega}, \mathbf{b}, \sigma_b) = \pi_0(\boldsymbol{\beta})\pi_1(\boldsymbol{\psi})\pi_2(\boldsymbol{\rho})\pi_3(\alpha)\pi_4(\phi)\pi_5(\mathbf{b}|\sigma_b)\pi_6(\sigma_b)$ and $\pi_j(.), j = 0, \ldots, 6$ denote the prior/hyperprior distributions on the model parameters as described above. The full conditional distributions necessary for the MCMC algorithm (combination of Gibbs sampling and Metropolis-within-Gibbs) in the AugGPD-RE model are as follows:

- The full conditional density for $\boldsymbol{\psi}|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\psi})}$, $\pi\left(\boldsymbol{\psi}|\mathbf{y}, \mathbf{b}, \sigma_b, \boldsymbol{\Omega}_{(-\boldsymbol{\psi})}\right)$ is proportional to

  $$\exp\left\{-\tfrac{1}{2}\boldsymbol{\psi}^\top \boldsymbol{\Sigma}_\psi^{-1} \boldsymbol{\psi}\right\} \prod_{i=1}^{n} \prod_{j=1}^{n_i} {p_0}_{ij}^{I_{y_{ij}=0}} (1 - {p_0}_{ij} - {p_1}_{ij})^{I_{y_{ij} \in (0,1)}}.$$

- The full conditional density for $\boldsymbol{\rho}|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\rho})}$, $\pi\left(\boldsymbol{\rho}|\mathbf{y}, \mathbf{b}, \sigma_b, \boldsymbol{\Omega}_{(-\boldsymbol{\rho})}\right)$ is proportional to

  $$\exp\left\{-\tfrac{1}{2}\boldsymbol{\rho}^\top \boldsymbol{\Sigma}_\rho^{-1} \boldsymbol{\rho}\right\} \prod_{i=1}^{n} \prod_{j=1}^{n_i} {p_1}_{ij}^{I_{y_{ij}=1}} (1 - {p_0}_{ij} - {p_1}_{ij})^{I_{y_{ij} \in (0,1)}}.$$

- The full conditional density for $\boldsymbol{\beta}|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\beta})}$, $\pi\left(\boldsymbol{\beta}|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\boldsymbol{\beta})}\right)$ is proportional to

  $$\exp\left\{-\tfrac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_\beta^{-1} \boldsymbol{\beta}\right\} \prod_{i=1}^{n} \prod_{j=1}^{n_i} g(y_{ij}; \alpha, \mu_{ij}, \phi)^{I_{y_{ij} \in (0,1)}},$$ with $g(y_{ij}; \alpha, \mu_{ij}, \phi)$ given by (4.3.5).

- The full conditional density for $\phi|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\phi)}$, $\pi(\phi|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\phi)})$ is proportional to $\pi(\phi) \prod_{i=1}^{n} \prod_{j=1}^{n_i} g(y_{ij}; \alpha, \mu_{ij}, \phi)^{I_{y_{ij} \in (0,1)}}$.
- The full conditional density for $\alpha|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\alpha)}$, $\pi(\alpha|\mathbf{y}, \mathbf{b}, \sigma_b, \Omega_{(-\alpha)})$ is proportional to $\prod_{i=1}^{n} \prod_{j=1}^{n_i} g(y_{ij}; \alpha, \mu_{ij}, \phi)^{I_{y_{ij} \in (0,1)}} I_{\alpha \in [0,1]}$.
- The full conditional density for $\mathbf{b}_i|\mathbf{y}, \sigma_b, \Omega$, $\pi(\mathbf{b}_i|\mathbf{y}, \sigma_b, \Omega)$ is proportional to $\exp\left\{-\sum_{k=1}^{q} \frac{1}{2\sigma_b^2} b_{ik}^2\right\} \prod_{j=1}^{n_i} g(y_{ij}; \alpha, \mu_{ij}, \phi)^{I_{y_{ij} \in (0,1)}}$ with b_{ik} the k-th element of $\mathbf{b}_i = (b_{i1}, \dots b_{iq})^\top$.
- The full conditional density for $\sigma_b|\mathbf{y}, \mathbf{b}, \Omega$, $\pi(\sigma_b|\mathbf{y}, \mathbf{b}, \Omega)$ is proportional to $\exp\left\{-\frac{1}{2\sigma_b^2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} b_{ij}^2 \}\right\} I_{\sigma_b \in (0, c_1)}$.

For specific densities of the GPD class, the full conditionals for the beta, BRe and simplex models are presented in Appendix B. For computational simplicity, we avoid the multivariate prior specifications for $\boldsymbol{\beta}$, $\boldsymbol{\psi}$ and $\boldsymbol{\rho}$ (multivariate zero mean vector with inverted-Wishart covariance) and instead assign simple i.i.d $\text{Normal}(0, \text{Variance} = 100)$ priors on the elements of these vectors, which centers the 'odds-ratio' type inference at 1 with a sufficiently wide 95% interval. When p_0 and p_1 represent constant proportions for the whole data, we allocate the Dirichlet prior with hyperparameter $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^\top$ for the probability vector $(p_0, p_1, 1 - p_0 - p_1)^\top$, with $\alpha_s \sim \text{Gamma}(1, 0.01)$, $s = 1, 2, 3$. After discarding the first 50000 burn-in samples, we used 50000 more samples (with a spacing of 10) from 2 independent chains with widely dispersed starting values for posterior summaries. Convergence was monitored via MCMC trace plots, autocorrelation plots and the Brooks-Gelman-Rubin $\hat{\text{R}}$ statistics. Associated R code is available on request from the corresponding author.

4.3.4 Bayesian model selection and influence diagnostics

For model selection, we use the conditional predictive ordinate (CPO) and the log pseudo-marginal likelihood (LPML) statistic (Carlin and Louis, 2008), derived from the posterior predictive distribution (ppd). Larger values of LPML indicate better fit. Computing CPO via the harmonic mean identity can lead to instability(Raftery et al., 2007). Hence, we consider a more pragmatic route and compute the CPO (and LPML) statistics using 500 non-overlapping blocks of the Markov chain, each of size 2000, post-convergence and report the expected LPML computed over the 500 blocks. In addition, we also apply the expected AIC (EAIC), expected BIC (EBIC) (Carlin and Louis, 2008) and the DIC_3 (Celeux et al., 2006) criteria. The DIC_3 was used as an alternative to the usual DIC (Spiegelhalter et al., 2002) because of the ease of computation directly from the MCMC output, and also due to the mixture modeling framework. All these criteria abide by the 'lower is better' law.

In addition, as a direct byproduct from the MCMC output, we develop some influence diagnostic measures to assess outlier effects on the fixed effects parameters based on case-deletion statistics (Cook and Weisberg, 1982), and the q-divergence measures (Csisz et al., 1967; Weiss, 1996) between posterior distributions. We consider three choices of these divergences, namely, the Kullback-Leibler (KL) divergence, the J-distance (symmetric version of the KL divergence), and the L_1-distance. We use the calibration methodPeng and Dey (1995) to obtain the cut-off values as 0.90, 0.83 and 1.32 for the L_1, KL and J-distances, respectively.

4.4 Data analysis and findings

The motivating PrD dataset assessed the PrD status of Gullah-speaking African-Americans with Type-2 diabetes via a detailed questionnaire focusing on demographics, social, medical and dental history. The dataset contain measurements on 28 teeth (considered full dentition, excluding the 4 third-molars) from 290 subjects, recording proportion of diseased tooth-sites (with CAL value $\geq$ 3mm) per tooth type as the response for each subject. Hence, this clustered data framework has 4 observations (corresponding to the 4 tooth-types) for each subject. If a tooth is missing, it was considered 'missing due to PrD' where all sites for that tooth contributed to the diseased category. Subject-level covariables in the dataset include gender (0=male,1= female), age of subject at examination (in years, ranging from 26 to 87 years), glycosylated hemoglobin (HbA1c) status indicator (0=controlled,$<$ 7%; 1=uncontrolled,$\geq$ 7%) and smoking status (0=non-smoker,1=smoker). We also considered a tooth-level variable representing each of the four tooth types, with 'canine' as the baseline.

From Figure 1 (left panel), the data are continuous on [0,1], with non-negligible proportions of of 0's (114, 9.8%) and 1's (94, 8.1%). Avoiding transformation, modeling via one of the members of the GPD class might not be feasible. Hence, we proceed using the AugGPD-RE model, adjusted for subject-level clustering. From Equations (4.3.1), (4.3.2) and (4.3.3), we have

$$\begin{aligned} \text{logit}(\mu_{ij}) &= \mathbf{X}_{ij}^{\top}\boldsymbol{\beta} + b_i, \qquad (4.4.1)\\ \text{logit}(p_{0ij}) &= \boldsymbol{W_0}_{ij}^{\top}\boldsymbol{\psi},\\ \text{logit}(p_{1ij}) &= \boldsymbol{W_1}_{ij}^{\top}\boldsymbol{\rho}, \end{aligned}$$

where $\mathbf{X}_{ij} = (1, \text{Gender}_{ij}, \text{Age}_{ij}, \text{HbA1c}_{ij}, \text{Smoker}_{ij}, \text{Incisor}_{ij}, \text{Premolar}_{ij}, \text{Molar}_{ij})^{\top}$, $\mathbf{X}_{ij} = \boldsymbol{W_0}_{ij} = \boldsymbol{W_1}_{ij}$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_7)^{\top}$, $\boldsymbol{\psi} = (\psi_0, \dots, \psi_7)^{\top}$ and $\boldsymbol{\rho} = (\rho_0, \dots, \rho_7)^{\top}$ are the vectors of regression parameters, and b_i is the subject-level random effect. The examination age was standardized (subtracting the mean and dividing by its standard deviation) to achieve better convergence. We have 6 competing models, varying with the densities in the GPD class and the regression over p_0 and p_1, as follows:

Model 1 $Y_{ij} \sim \text{ZOAS-RE}(\mu_{ij}, \phi, p_{0ij}, p_{1ij})$.
Model 1a $Y_{ij} \sim \text{ZOAS-RE}(\mu_{ij}, \phi, p_0, p_1)$.
Model 2 $Y_{ij} \sim \text{ZOAB-RE}(\mu_{ij}, \phi, p_{0ij}, p_{1ij})$.
Model 2a $Y_{ij} \sim \text{ZOAB-RE}(\mu_{ij}, \phi, p_0, p_1)$.
Model 3 $Y_{ij} \sim \text{ZOABRe-RE}(\alpha, \mu_{ij}, \phi, p_{0ij}, p_{1ij})$.
Model 3a $Y_{ij} \sim \text{ZOABRe-RE}(\alpha, \mu_{ij}, \phi, p_0, p_1)$.

Note that the parameter α is specific to the ZOABRe model only. In addition, we also fit the LS-simplex model (or **Model 4**) by transforming the response from y to y' via the Lemon-squeezer (LS) transformation(Smithson and Verkuilen, 2006) given by $y' = [y(N-1)+1/2]/N$, where N is the number total of observations, with the regression on μ as (4.4.1). Although models 1, 1a, 2, 2a, 3 and 3a can be compared using standard model choice criteria described in Subsection 4.3.4 because they fit the same dataset, this is not the case for the LS-simplex model which fits a transformed dataset. Thus, we assess its fit visually via the empirical cumulative distribution functions (ecdfs) of the fitted values. Table 4.1 presents the DIC$_3$, LPML, EAIC and EBIC values for the 6 competing models. Notice that Model 1 (ZOAS-RE model) provides the best fit uniformly across all criteria. Also, the fit for models with constant p_0 and p_1 are worser than the corresponding ones with regression on p_0 and p_1. The right panel of Figure 4.1 clear tells us that the ecdf from the fitted values using Model 1 represent the true data much closely as compared to Model 4. Hence, we select Model 1 as our best model and proceed with inference.

Criterion	Model					
	1	1a	2	2a	3	3a
DIC$_3$	**915.3**	1165.3	993.0	1243.5	1001.3	1253.4
LPML	**-461.1**	-584.1	-500.5	-623.7	-503.8	-627.8
EAIC	**917.8**	1154.9	992.7	1231.0	967.4	1210.4
EBIC	**1047.2**	1210.5	1124.2	1286.6	1103.9	1281.2

Table 4.1: Model comparison using DIC$_3$, LPML, EAIC and EBIC criteria.

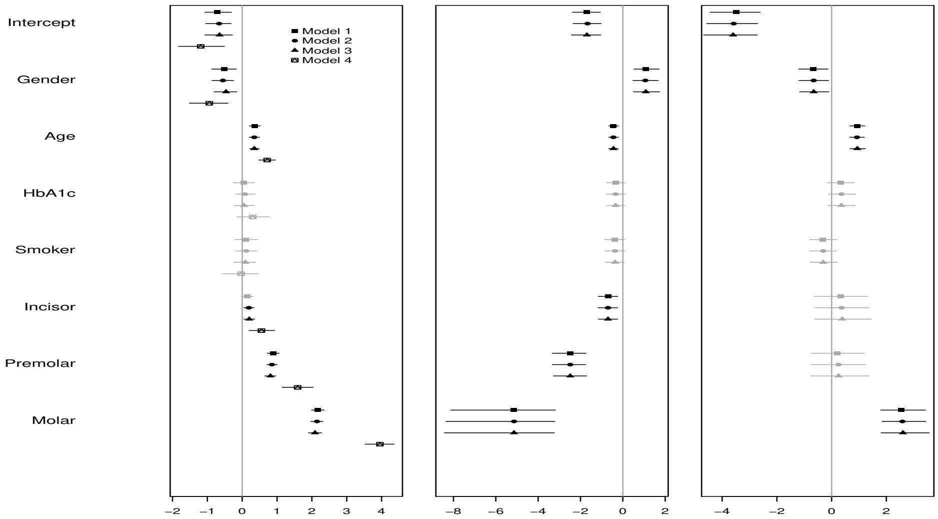

Figure 4.3: Posterior mean and 95% credible intervals (CI) of parameter estimates from Models 1-4 for μ (left pannel), for p_0 (middle pannel) and for p_1 (right pannel). CIs that include zero are gray, those that does not include zero are black.

Plots of the means of the posterior parameter estimates and their 95% CIs for the regression onto μ (left panel), p_0 (middle panel) and p_1 (right panel) for Models 1-4 are presented in Figure 4.3. We do not report the estimates from the models that consider p_0 and p_1 as constants (i.e., Models 1a, 2a, and 3a). In this Figure, the gray intervals contain zero (non-significant covariates), while the black intervals do not include zero and are considered significant at 5% level. From the left panel (regression onto μ_{ij}), the covariates gender, age and tooth-types significantly explain the proportion responses mostly for Models 1-4, with the exception of Incisor for Model 1 where it is non-significant. Parameter interpretation can be expressed in terms of its effect directly on μ_{ij}, specifically $\frac{\mu_{ij}}{1-\mu_{ij}}$, conditional on the set of other covariates and REs Galvis et al. (2014). Here, μ_{ij} is the 'expected proportion of diseased sites, and $1 - \mu_{ij}$ is the complement, i.e., the 'expected remaining proportion to being completely diseased', both conditional on μ_{ij} not being zero or one. These results are interpreted in terms of the number of times the ratio is higher/lower 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 example, for age (a strong predictor of PrD), this ratio is 1.43 ($\exp(0.36) = 1.43$, 95% CI=$[1.23, 1.66]$) times higher for every unit increase in Age. For Gender, this ratio is 40% lower for males as compared to females, which might be influenced by the lower participation of males common in this population (Johnson-Spruill et al., 2009). Similarly, this ratio is 8.7 times higher for molars as compared to the canines (the baseline), which confirms that the posteriorly placed molars typically experience a higher PrD status than the anterior canines. From the plots in the middle and right panels of Figure 4.3, we identify gender, age and tooth-types to be significant in explaining absence of PrD, while gender, age and molar significantly explaining the completely diseased category. Once again, we have similar odds-ratio explanation as earlier. For example, the odds of a tooth type free of PrD are 3 times greater for men than for women, while the odds of a completely diseased molar are about 13 times than of a (baseline) canine. Rest of the parameters can be interpreted similarly.

The mean estimates (standard deviations) of ϕ from Models 1-4 are 0.14 (0.007), 7.6 (0.43), 10.6 (1.56) and 0.002 (< 0.0001), and of σ_b^2 are 1.3 (0.13), 1.2 (0.13), 1.2 (0.13) and 2.6 (0.34), respectively. Due to parametrization involved, these estimates of ϕ are not comparable across Models 1-3. However, the effect of the LS transformation is evident while comparing the estimates between Models 1 and 4. Additionally, the estimates of σ_b^2 reveal that the transformation in Model 4 leads to a higher (estimated) variance of the response Y than the Models 1-3.

The adequacy of the logit link is assessed via plots of the linear predictor versus the predicted probability (Hatfield et al., 2012) as depicted in the Figure in Appendix C. Considering $\text{logit}^{-1}(\mu_{ij})$ from Model 1, we divided it into 10 intervals containing roughly an equal number of observations, and plot the distribution of the inverse-logit transformed linear predictors (denoted by the black box-plots) that represents 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. There seem to be no evidence of model misspecification, i.e., the shapes of the fitted and observed trends are similar, as revealed from Figure C in the Appendix.

In addition, we conduct sensitivity analysis on the prior assumptions for the random effects precision ($1/\sigma_b^2$) and the fixed effects precision parameters on $\boldsymbol{\beta}$ by changing one parameter at a time and refitting Model 1, as in (Galvis et al., 2014). In particular, we allowed $\sigma_b \sim \text{Uniform}(0, k)$, where $k \in \{10, 50\}$, and also the typical Inverse-gamma choice on the precision $1/\sigma_b^2 \sim \text{Gamma}(k, k)$, where $k \in \{0.001, 0.1\}$. We also chose the normal precision on the fixed effects to be 0.1, 0.25 (which reflects an odds-ratio in between e^{-4} to e^4) and 0.001. There were slight changes observed in parameter estimates and model comparison values, however, that did not change our conclusions regarding the best model, inference (and sign) of the fixed-effects, and the influential observations.

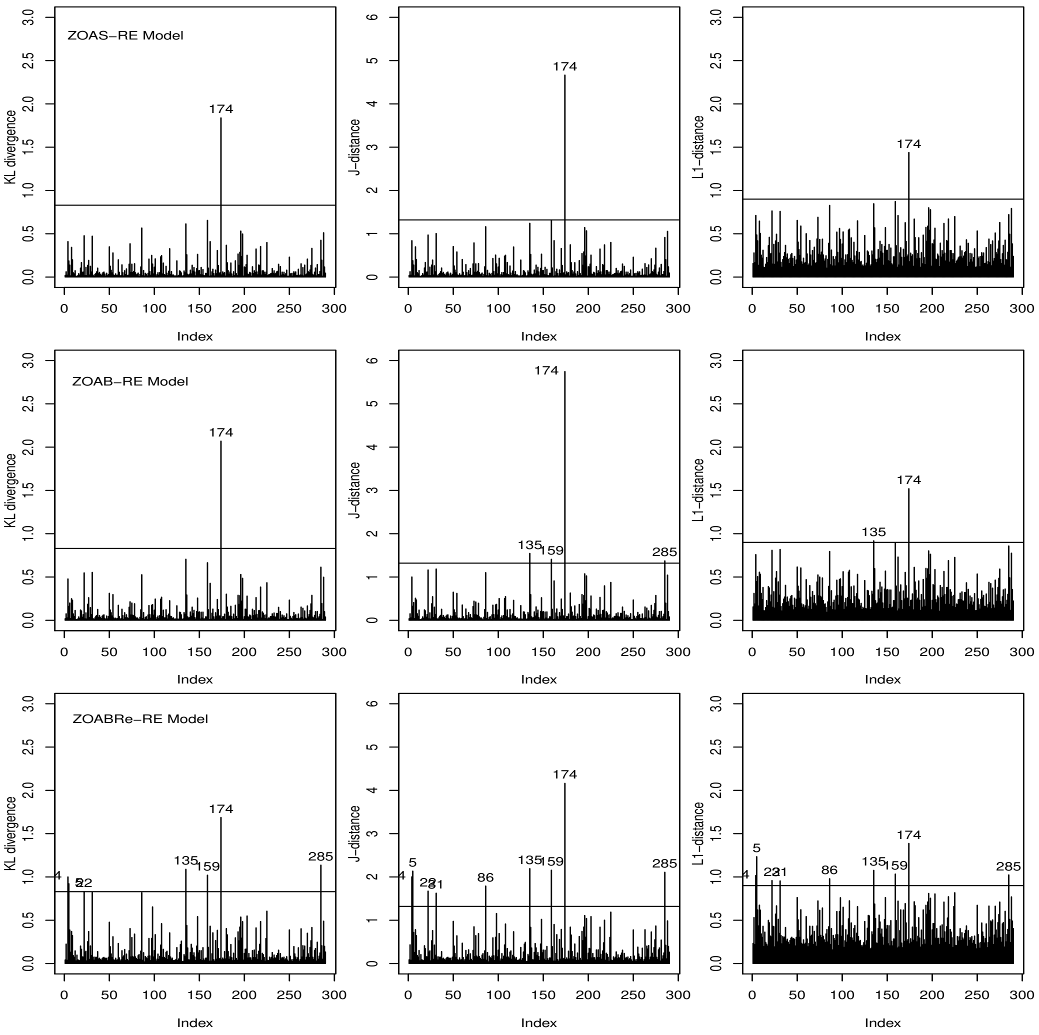

Figure 4.4: K-L, J and L1 divergences from the ZOAS-RE (upper panel), ZOAB-RE (middle panel) and ZOABRe-RE (lower panel) models for the PrD dataset.

Finally, we detect outlying observations via the q-divergence measures for the augmented models using the cut-offs described in Subsection 4.3.4. These plots are presented in Figure 4.4, where the upper, middle and lower panels represent the ZOAS-RE, ZOAB-RE and ZOABRe-RE models, respectively. Interestingly, we find that the ZOABRe-RE model produces several outlying observations exceeding the threshold, whereas the best-fitting model

(ZOAS-RE) produces only one such observation (subject id # 174). To quantify the impact of this observation, we refit the model by removing it. The covariate 'Molar' in the regression onto p_{0ij} is impacted by this observation, perhaps due to this subject is free of PrD for all tooth types. However, the parameter significance and sign of the coefficients remained the same. Henceforth, we stick to the estimates obtained from fitting Model 1 to the full data, without removing this particular subject.

4.5 Simulation studies

In order to assess the finite sample performance of the class of AugGPD-RE mixed regression models, we conduct two simulation studies. First (Scheme 1), we assess the impact of model misspecification on the parameters for the ZOAS-RE, ZOAB-RE and ZOABRe-RE models when the data in (0,1) are generated from a logistic normal model (Atchison and Shen, 1980). Next (Scheme 2), we analyze the impact of the LS transformation on the parameter estimates in presence of various proportions of zeros and ones. In both studies, we generate data with various sample sizes, and compare the mean squared error (MSE), absolute relative bias (Abs.RelBias), and coverage probability (CP) of the regression parameters across the various models.

Initially, we generate y_{ij} for both schemes and sample sizes $n = 50, 100, 150, 200$ as $y_{ij} = \text{logit}^{-1}(T_{ij})$, $i = 1, \ldots, n$ (the number of subjects), $j = 1, \ldots, 5$ (indicating cluster of size 5 for each subject), with $T_{ij} \sim \text{Normal}(\mu_{ij}, 1)$ and the location parameter μ_{ij} modeled as $\mu_{ij} = \beta_0 + \beta_1 x_{ij} + b_i$, with $b_i \sim N(0, \sigma_b^2)$. The explanatory variables x_{ij} are generated as independent draws from a Uniform$(0, 1)$, with the regression parameters fixed at $\beta_0 = -0.5$, and $\beta_1 = 0.5$, variance component $\sigma^2 = 2$, and constant proportions $p_0 = 0.1$ and $p_1 = 0.1$. Thus, $y_{ij} \in (0, 1)$ are draws from a logistic-normal model. Finally, via multinomial sampling, we allocate the 0's, 1's, and the $y_{ij} \in (0, 1)$ with probabilities p_0, p_1 and $1-p_0-p_1$ respectively. No regression onto p_0 and p_1 are considered.

After simulating 200 such datasets, we fitted the ZOAS-RE, ZOAB-RE and ZOABRe-RE models with similar prior choices as in the data analysis. With our parameter vector $\boldsymbol{\theta} = (\beta_0, \beta_1, p_0, p_1, \sigma_b^2)$, and θ_s being an element of $\boldsymbol{\theta}$, we calculate the MSE as $\text{MSE}(\hat{\theta}_s) = \frac{1}{200}\sum_{i=1}^{200}(\hat{\theta}_{is} - \theta_s)^2$, the absolute relative bias as Abs.RelBias $(\hat{\theta}_s) = \frac{1}{200}\sum_{i=1}^{200}|\frac{\hat{\theta}_{is}}{\theta_s} - 1|$, and the 95% coverage probability (CP) as $\text{CP}(\hat{\theta}_s) = \frac{1}{200}\sum_{i=1}^{200} I(\theta_s \in [\hat{\theta}_{s,LCL}, \hat{\theta}_{s,UCL}])$, 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 bounds of the 95% limits of the CIs, respectively. The results from this study for varying sample sizes are presented in Figure 4.5 and Table 1 (Appendix D). Figure 4.5 presents a visual comparison of the models (bold line for the ZOAS-RE model, dashed line for the ZOAB-RE model and dotted line for the ZOABRe-RE model) for β_0 (upper panel) and β_1 (lower panel). For the sake of brevity, we do not produce plots for p_0, p_1 and σ_b^2. We observe that the Abs.RelBias of both β_0, β_1 and σ_b^2 are much smaller for the ZOAS-RE model as compared to the ZOAB-RE model and the ZOABRe-RE models, while those for p_0 and p_1 are comparable. The MSEs of the parameters other than σ_b^2 are comparable. For σ_b^2, the ZOAS-RE performs better (MSE is lower) than the other two. CP remains higher for the ZOAS-RE as compared to the other two models across all

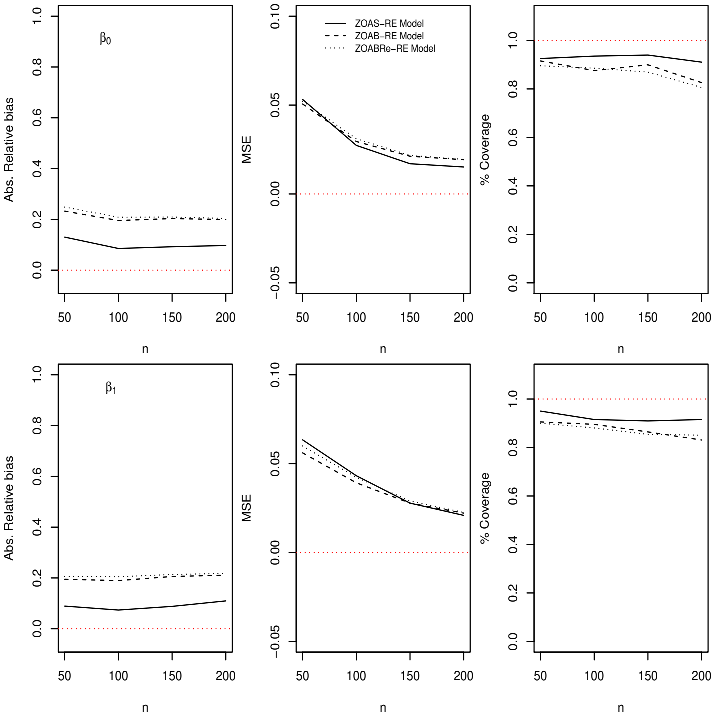

Figure 4.5: Absolute relative bias, MSE and coverage probability of β_0 and β_1 after fitting ZOAS-RE (continuous), ZOAB-RE (dashed) and ZOABRe-RE (dotted) models.

parameters. Interestingly, for σ_b^2, the CP is estimated close to zero for higher n ($n = 150, 200$)

In Scheme 2, we compare the performance of the ZOAS-RE and LS-simplex models for three scenarios of p_0 and p_1, namely (a): $p_0 = p_1 = 1\%$, (b) $p_0 = 3\%, p_1 = 5\%$, and (c) $p_0 = 10\%, p_1 = 8\%$ (that represents the real data). Figure 4.6 present the plots for MSE, Abs.RelBias and CP. The ZOAS-RE outperforms the LS-simplex model with lower MSE and Abs.RelBias, and higher CP across all scenarios, with the performance of the simplex model getting worser with increase in the proportion of 0's and 1's.

4.6 Conclusions

Motivated by the presence of extreme proportion responses, we develop a class of (parametric) augmented proportion density models under a Bayesian framework, and demonstrate its application to a PrD dataset. As a byproduct of the MCMC output, we also develop tools for outlier detection using results from q-divergence measures. Both simulation and real data analysis reveal the importance of utilizing an appropriate theoretical model over ad hoc data transformations.

Note that in our model development, we regress the covariates onto μ_{ij} as in Definition 2. For a direct interpretation of the covariate effect on the response Y, one might consider regressing onto δ_{ij} (the conditional expectation of the true AugGPD response) via some link functions. However, on applying this to our dataset, we experienced problems with MCMC

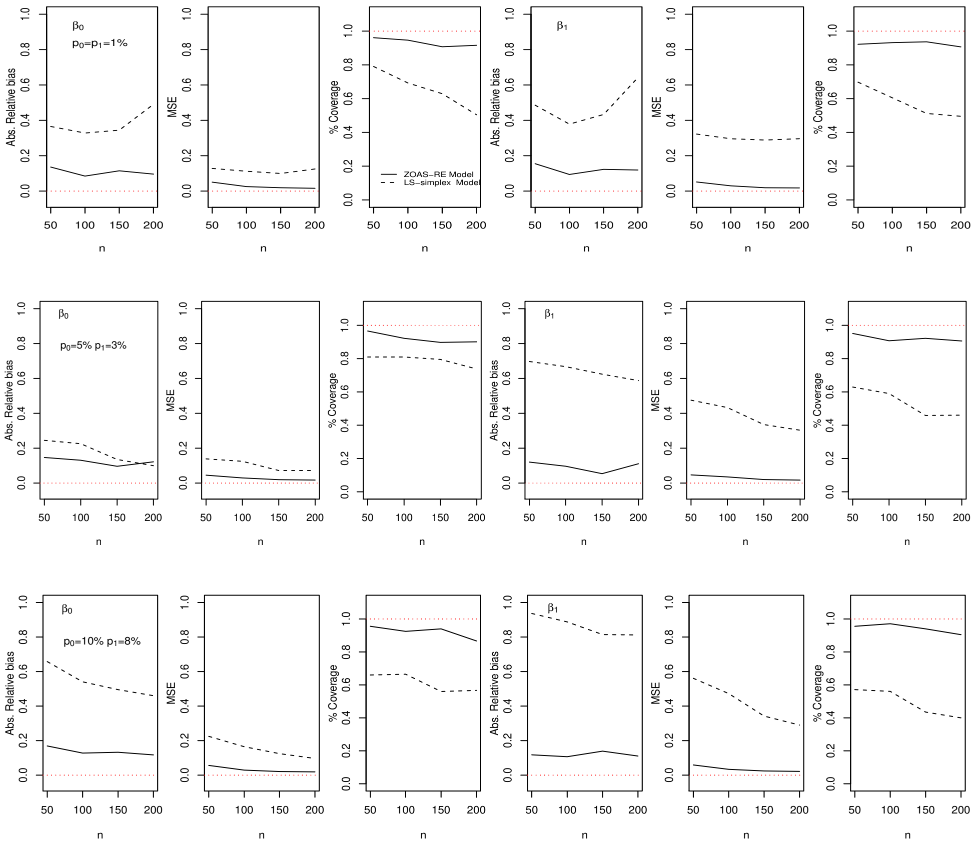

Figure 4.6: Absolute relative bias, MSE and coverage probability of β_0 and β_1 after fitting ZOAS-RE (continuous)and LS-simplex (dashed) models, for $p_0 = p_1 = 1\%$ (upper panel), $p_0 = 5\%$, $p_1 = 3\%$ (middle panel) and $p_0 = 10\%$, $p_1 = 8\%$ (lower panel).

convergence. Hence, we did not pursue it any further, although it may be appropriate for other datasets.

The current clustered setup can be extended to a longitudinal, or a clustered-longitudinal framework (often found in dental clinical trials). In addition, the current development explores a simple parametric framework with ease in implementation. Certainly, the shape of the proportion data can also be adequately captured via some (flexible) nonparametric specification of the density. However, the Bayesian implementation may not be automatic, and would require developing customized MCMC algorithms. All these remain viable components of future research.

APPENDIX

APPENDIX A: Some densities in the GPD class

- *The beta distribution*

The density of a r.v Y following the beta distribution with mean μ and precision parameter

ϕ is given by

$$f(y|\mu,\phi)=\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},\tag{A1}$$

with $0<E[Y]=\mu<1$, $\mathrm{Var}(Y)=\frac{\mu(1-\mu)}{1+\phi}$ and $\phi>0$.
• *The simplex distribution*
A r.v Y follows a simplex distribution with parameters μ and ϕ if its pdf is given by

$$f(y|\mu,\phi)=\frac{\sqrt{\phi}}{\left(\pi\left[y(1-y)\right]^3\right)^{1/2}}\exp\left\{-\phi\frac{(y-\mu)^2}{2y(1-y)\mu^2(1-\mu)^2}\right\},\tag{A2}$$

with $0<E[Y]=\mu<1$ and $\phi>0$.
• *The beta rectangular*
A r.v Y is distribuited according to beta rectangular distribution with parameters η, λ and ϕ if its pdf is given by

$$f(y|\eta,\lambda,\phi)=\eta+(1-\eta)\frac{\Gamma(\phi)}{\Gamma(\lambda\phi)\Gamma((1-\lambda)\phi)}y^{\lambda\phi-1}(1-y)^{(1-\lambda)\phi-1},\tag{A3}$$

with $0\leq\eta\leq1$, $0<\lambda<1$, $\phi>0$, $E[Y]=\eta/2+(1-\eta)\lambda$ and $\mathrm{Var}(Y)=\frac{\lambda(1-\lambda)}{1+\phi}(1-\eta)(1+\eta(1+\phi))+\frac{\eta}{12}(4-3\eta)$.

APPENDIX B: Full conditional distributions from models ZOAS-RE, ZOAB-RE and ZOABRe-RE in the augmented GPD class

The full conditional distributions of the parameters $\boldsymbol{\psi}$, $\boldsymbol{\rho}$ and σ_b necessary for the MCMC algorithm in the three models above remain equal to presented for the augmented-GPD class. For the other parameters, the full conditional distributions are obtained for every model as follows.

ZOAS-RE model

- The full conditional density for $\boldsymbol{\beta}|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\boldsymbol{\beta})}$, $\pi\left(\boldsymbol{\beta}|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\boldsymbol{\beta})}\right)$ is proportional to
$$\exp\left\{-\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)-\phi\sum_{i=1}^{n}\sum_{j=1}^{n_i}\frac{\left[y_{ij}-(1-y_{ij})A_{ij}\right]^2(1+A_{ij})^2}{2y_{ij}(1-y_{ij})A_{ij}^2}I_{y_{ij}\in(0,1)}\right\},$$
where $A_{ij}=\exp\{\mathbf{X}_{ij}^\top\boldsymbol{\beta}+\mathbf{Z}_{ij}^\top\mathbf{b}_i\}$.

- The full conditional density for $\phi|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\phi)}$, $\pi(\phi|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\phi)})$ is a left truncated gamma with left truncation point a^{-2}. That is $\phi|\mathbf{y},\mathbf{b},\sigma_b,\boldsymbol{\Omega}_{(-\phi)}\sim TGamma^{-}(a^{-2},(n-1)/2,\sum_{i=1}^{n}\sum_{j=1}^{n_i}a_3(y_{ij},\mu_{ij}))$ where $a_3(y_{ij},\mu_{ij})=\frac{(y_{ij}-\mu_{ij})^2}{2y_{ij}(1-y_{ij})\mu_{ij}^2(1-\mu_{ij})^2}$ and $\mu_{ij}=\frac{A_{ij}}{1+A_{ij}}$.

- The full conditional density for $\mathbf{b}_i|\mathbf{y},\sigma_b,\boldsymbol{\Omega}$, $\pi(\mathbf{b}_i|\mathbf{y},\sigma_b,\boldsymbol{\Omega})$ is proportional to
$\exp\left\{\frac{-1}{2\sigma_b^2}\sum_{k=1}^{q}b_{ik}^2-\phi\sum_{j=1}^{n_i}\frac{[y_{ij}-(1-y_{ij})A_{ij}]^2(1+A_{ij})^2}{2y_{ij}(1-y_{ij})A_{ij}^2}I_{y_{ij}\in(0,1)}\right\}I_{\{b_{ik}\in\mathbb{R}\}}.$

ZOAB-RE model

- The full conditional density for $\boldsymbol{\beta}|\mathbf{y}, \mathbf{b}, \sigma_b, \boldsymbol{\Omega}_{(-\boldsymbol{\beta})}$, $\pi\left(\boldsymbol{\beta}|\mathbf{y}, \mathbf{b}, \sigma_b, \boldsymbol{\Omega}_{(-\boldsymbol{\beta})}\right)$ is proportional to

$$\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(\mu_{ij}\phi\log\frac{y_{ij}}{1-y_{ij}} - B_{ij}\right) I_{y_{ij}\in(0,1)}\right\},$$

  where $B_{ij} = \log\Gamma(\mu_{ij}\phi) + \log[\Gamma(1-\mu_{ij})\phi]$, $\mu_{ij} = \frac{A_{ij}}{1+A_{ij}}$.

- The full conditional density for $\phi|\mathbf{y}, \mathbf{b}, \sigma_b^2, \boldsymbol{\Omega}_{(-\phi)}$, $\pi(\phi|\mathbf{y}, \mathbf{b}, \sigma_b^2, \boldsymbol{\Omega}_{(-\phi)})$ is proportional to

$$\phi^{a-1}\exp\left\{-\phi\left(c - \sum_{i=1}^{n}\sum_{j=1}^{n_i} C_{ij} I_{y_{ij}\in(0,1)}\right)\right\},$$

  where $C_{ij} = \mu_{ij}\log\frac{y_{ij}}{1-y_{ij}} + (1-\mu_{ij})\log(1-y_{ij}) + \log(\phi) - B_{ij}$ and $\phi > 0$.

- The full conditional density for $\mathbf{b}_i|\mathbf{y}, \sigma_b^2, \boldsymbol{\Omega}$, $\pi(\mathbf{b}_i|\mathbf{y}, \sigma_b^2, \boldsymbol{\Omega})$ is proportional to

$$\exp\left\{\frac{-1}{2\sigma_b^2}\sum_{k=1}^{q} b_{ik}^2 - \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(\mu_{ij}\phi\log\frac{1-y_{ij}}{y_{ij}} - B_{ij}\right) I_{y_{ij}\in(0,1)}\right\},$$

  with $b_{ik} \in \mathbb{R}$.

ZOABRe-RE model

- The full conditional density for $\boldsymbol{\beta}|\mathbf{y}, \mathbf{b}, \sigma_b^2, \boldsymbol{\Omega}_{(-\boldsymbol{\beta})}$, $\pi\left(\boldsymbol{\beta}|\mathbf{y}, \mathbf{b}, \sigma_b^2, \boldsymbol{\Omega}_{(-\boldsymbol{\beta})}\right)$ is proportional to
  $\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + \sum_{i=1}^{n}\sum_{j=1}^{n_i} I_{\{y_{ij}\in(0,1)\}}\log\left[\eta_{ij} + (1-\eta_{ij})M_{ij}\right]\right\}$,
  where $M_{ij} = \frac{\Gamma(\phi)}{\Gamma(\lambda_{ij}\phi)\Gamma((1-\lambda_{ij})\phi)} y_{ij}^{\lambda_{ij}\phi-1}(1-y_{ij})^{(1-\lambda_{ij})\phi-1}$, $\eta_{ij} = \alpha(1-2|\mu_{ij}-\frac{1}{2}|)$, $\lambda_{ij} = \frac{\mu_{ij}-\frac{\eta_{ij}}{2}}{1-\eta_{ij}}$ and $\mu_{ij} = \frac{A_{ij}}{1+A_{ij}}$.

- The full conditional density for $\phi|\mathbf{y}, \mathbf{b}, \sigma_b^2, \boldsymbol{\Omega}_{(-\phi)}$, $\pi(\phi|\mathbf{y}, \mathbf{b}, \sigma_b^2, \boldsymbol{\Omega}_{(-\phi)})$ is proportional to
  $\phi^{a-1}\exp\left\{-\phi c + \sum_{i=1}^{n}\sum_{j=1}^{n^i} I_{\{y_{ij}\in(0,1)\}}\log\left[\eta_{ij} + (1-\eta_{ij})M_{ij}\right]\right\}$,
  with $\phi > 0$.

- The full conditional density for $\mathbf{b}_i|\mathbf{y}, \sigma_b^2, \boldsymbol{\Omega}$, $\pi(\mathbf{b}_i|\mathbf{y}, \sigma_b^2, \boldsymbol{\Omega})$ is proportional to
  $\exp\left\{\frac{-1}{2\sigma_b^2}\sum_{k=1}^{n_i} b_{ik}^2 + \sum_{i=1}^{n}\sum_{j=1}^{n_i} I_{\{y_{ij}\in(0,1)\}}\log\left[\eta_{ij} + (1-\eta_{ij})M_{ij}\right]\right\}$,
  with $b_{ij} \in \mathbb{R}$.

- The full conditional density for $\alpha|\mathbf{y}, \sigma_b^2, \boldsymbol{\Omega}$, $\pi(\alpha|\mathbf{y}, \sigma_b^2, \boldsymbol{\Omega})$ is proportional to
  $\exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{n_i} I_{\{y_{ij}\in(0,1)\}}\log\left[\eta_{ij} + (1-\eta_{ij})M_{ij}\right]\right\} I_{\{\alpha\in[0,1]\}}$.

APPENDIX C: Adequacy of the logit link

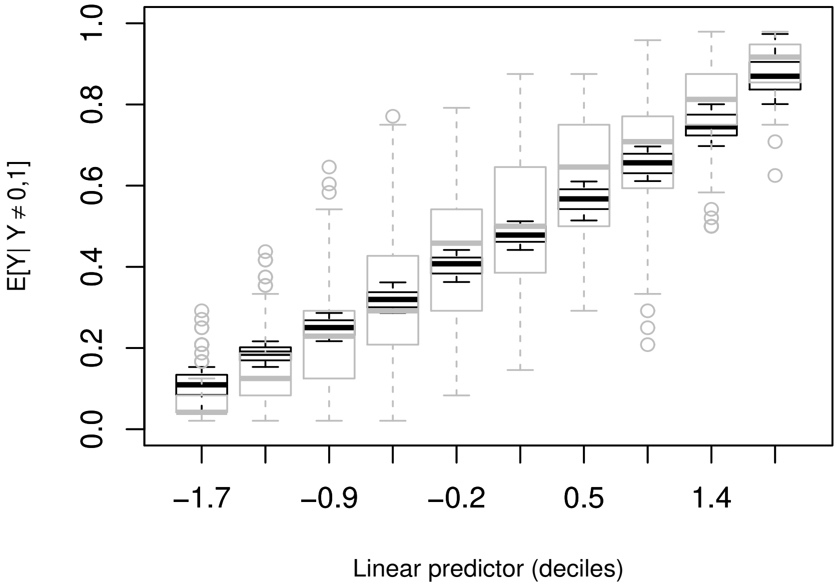

Figure 4.7: Observed and fitted relationship between the linear predictor and the (conditional) non-zero-one mean μ_{ij}. Modeled logit relationships are represented by black box-plots, while the empirical proportions by gray box-plots.

APPENDIX D: Simulation Results from Scheme 1

	ZOAS-RE model				ZOAB-RE model				ZOABRe-RE Model			
Parameter	$n = 50$	$n = 100$	$n = 150$	$n = 200$	$n = 50$	$n = 100$	$n = 150$	$n = 200$	$n = 50$	$n = 100$	$n = 150$	$n = 200$
				Abs.RelBias								
β_0	0.13	0.08	0.09	0.10	0.23	0.19	0.20	0.20	0.25	0.21	0.21	0.20
β_1	0.09	0.07	0.09	0.11	0.19	0.19	0.20	0.21	0.21	0.20	0.21	0.22
p_0	0.02	0.02	0.00	0.0001	0.02	0.02	0.0003	0.00058	0.02	0.022	0.0009	0.0002
p_1	0.05	0.005	0.01	0.01	0.05	0.005	0.01	0.01	0.05	0.005	0.011	0.01
σ_b^2	0.19	0.20	0.22	0.226	0.37	0.38	0.40	0.40	0.38	0.38	0.40	0.40
				MSE								
β_0	0.05	0.03	0.02	0.01	0.05	0.03	0.02	0.02	0.05	0.03	0.02	0.19
β_1	0.06	0.04	0.03	0.02	0.06	0.04	0.03	0.02	0.06	0.04	0.03	0.02
p_0	0.0004	0.0002	0.0001	8e-05	0.0004	0.0002	0.0001	8e-05	0.0004	0.0002	0.0001	8e-0
p_1	0.0004	0.0001	0.0001	8e-05	0.0004	0.0001	0.0001	8e-05	0.0004	0.0001	0.0001	8e-0
σ_b^2	0.28	0.23	0.24	0.24	0.63	0.62	0.66	0.66	0.65	0.63	0.66	0.66
				CP								
β_0	0.92	0.93	0.94	0.91	0.91	0.88	0.90	0.83	0.90	0.89	0.87	0.81
β_1	0.95	0.91	0.91	0.91	0.90	0.90	0.86	0.83	0.90	0.88	0.85	0.85
p_0	0.94	0.92	0.93	0.96	0.94	0.93	0.93	0.97	0.94	0.91	0.92	0.95
p_1	0.94	0.95	0.94	0.96	0.92	0.95	0.94	0.96	0.93	0.95	0.94	0.96
σ_b^2	0.82	0.68	0.50	0.35	0.49	0.16	0.01	0.05	0.49	0.15	0.01	0.0

Table 4.2: Absolute Relative bias (Abs.RelBias), mean squared error (MSE), and coverage probabilities (CP) of the the parameter estimates after fitting the ZOAS-RE, ZOAB-RE, and ZOABRe-RE models to simulated data for various sample sizes.

Chapter 5

Concluding remarks

5.1 Conclusions

Motivated by the presence of extreme proportion responses and the classical development of (Ospina and Ferrari, 2010), it was developed a class of (parametric) augmented proportion density models under a Bayesian perspective, and demonstrated its application to a Periodontal dataset. In this model development, there were regressed covariates onto μ_{ij}, p_{0ij}, p_{1ij}, leading to identifying covariates that are significant to explain disease-free, progressing with disease, and completely diseased tooth types. An alternative method (Smithson and Verkuilen, 2006) that transforms data was also studied and compared with the proposed models. Both simulation and real data analysis reveal the importance of utilizing an appropriate theoretical model over ad hoc data transformations. There were also developed tools for outlier detection using q-divergence measures, and quantified their effect on the posterior estimates of the model parameters.

Within the GPD class, the simplex density is more flexible than both its competitors. Hence, most likely the simplex regression (and its augmented counterpart) will outperform the beta and the Bre regressions for relatively non-smooth proportion data, such as, data with lots of spikes and structures, for support within $(0, 1)$ (and $[0, 1]$). However, it is recommended a pragmatic modeling approach by fitting these 3 parametric densities successively to any dataset, and choosing the best one via popular model selection techniques.

5.2 Other publications

- A Mixed-Effect Model for Positive Responses Augmented by Zeros. Mariana Rodrigues-Motta, Diana Milena Galvis Soto, Victor H. Lachos et al. Provisionally accepted paper in Statistics in Medicine. 2014.

- Bayesian semiparametric longitudinal data modeling using normal/independent densities. Luis M. Castro, Victor H. Lachos, Diana M. Galvis and Dipankar Bandyopadhyay. Aceito para publicação em Chapman & Hall/CRC Press. Edited volume in “Current Trends in Bayesian Methodology with Applications”. 2014.

5.3 Future research

It is of interest to investigate the presence of thick/heavy tails in the underlying ZOAB-RE, ZOAS-RE and ZOABr-RE models, and to model the random effect term b_i using robust alternatives (say, the t-density) over the normal density as in Figueroa-Zúniga et al. (2013). For periodontal dataset, the results were very similar using a t-density, and hence we did not consider it any further. The current analysis considers clustered cross-sectional periodontal proportion data. Often, these study subjects can be randomized to dental treatments and subsequent longitudinal follow-ups, leading to a clustered-longitudinal framework, where one might be interested in estimating the profiles (both overall, and subject-level) in the proportion of diseased surfaces for the four tooth types with time.

Certainly, the shape of the proportion data can also be adequately captured via some (flexible) nonparametric specification of the density. However, the Bayesian implementation may not be automatic, and would require developing customized MCMC algorithms. Also, it is possible to include a effect that model the spatial relation that exist in the application data as in Bandyopadhyay et al. (2011).

Bibliography

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* London: Chapman & Hall.

Atchison, J. and S. M. Shen (1980). Logistic-normal distributions: Some properties and uses. *Biometrika 67*(2), 261–272.

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

Bandyopadhyay, D., B. J. Reich, and E. H. Slate (2011). A spatial beta-binomial model for clustered count data on dental caries. *Statistical Methods in Medical Research 20*(2), 85–102.

Barndorff-Nielsen, O. E. and B. Jørgensen (1991). Some parametric models on the simplex. *Journal of Multivariate Analysis 39*(1), 106–116.

Bayes, C. L., J. L. Bazán, and C. García (2012). A new robust regression model for proportions. *Bayesian Analysis 7*(4), 841–866.

Branscum, A. J., W. O. Johnson, and M. C. Thurmond (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.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

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

Carrasco, J. M., S. L. Ferrari, and R. B. Arellano-Valle (2014). Errors-in-variables beta regression models. *Journal of Applied Statistics 41*(7), 1–18.

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

Cepeda-Cuervo, E. (2001). *Modeling variability in generalized linear models.* Ph. D. thesis, Mathematics Institute, Universidade Federal do Rio de Janeiro.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics 19*(1), 15–18.

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

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

Cowles, M. K. and B. P. Carlin (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., M.-H. Chen, and H. Chang (1997). Bayesian approach for nonlinear random effects models. *Biometrics*, 1239–1252.

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

Fernandes, J., C. Salinas, S. London, R. Wiegand, E. Hill, E. Slate, J. Grewal, P. Werner, J. Sanders, and M. Lopes-Virella (2006). Prevalence of periodontal disease in gullah african american diabetics. *Journal of Dental Research 85*(Special Issue A), 0997.

Ferrari, S. and F. Cribari-Neto (2004). Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics 31*(7), 799–815.

Figueroa-Zúniga, J. I., R. B. Arellano-Valle, and S. L. Ferrari (2013). Mixed beta regression: A Bayesian perspective. *Computational Statistics & Data Analysis 61*, 137–147.

Galvis, D. M., D. Bandyopadhyay, and V. H. Lachos (2014). Augmented mixed beta regression models for periodontal proportion data. *Statistics in Medicine 33*(21), 3759–3771.

Geisser, S. and W. F. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association 74*(365), 153–160.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian analysis 1*(3), 515–534.

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

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

Hahn, E. D. (2008). Mixture densities for project management activity times: A robust approach to PERT. *European Journal of Operational Research 188*(2), 450–459.

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

Jara, A., F. Quintana, and E. San Martín (2008). Linear mixed models with skew-elliptical distributions: A bayesian approach. *Computational Statistics & Data Analysis 52*(11), 5033–5045.

Johnson, N., S. Kotz, and N. Balakrishnan (1994). *Continuous Univariate Distributions, Vol. 2*. New York: John Wiley & Sons.

Johnson-Spruill, I., P. Hammond, B. Davis, Z. McGee, and D. Louden (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.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 127–162.

Jørgensen, B. (1997). *The Theory of Dispersion Models*, Volume 76. CRC Press.

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

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology 46*(1), 79–88.

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

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

Lachos, V. H., L. M. Castro, and D. K. Dey (2013). Bayesian inference in nonlinear mixed–effects models using normal independent distributions. *Computational Statistics & Data Analysis 64*, 237–252.

Lachos, V. H., D. K. Dey, and V. G. Cancho (2009). Robust linear mixed models with skew-normal independent distributions from a Bayesian perspective. *Journal of Statistical Planning and Inference 139*, 4098–4110.

Ligges, U., A. Thomas, D. Spiegelhalter, N. Best, D. Lunn, K. Rice, and S. Sturtz (2009). BRugs 0.5: OpenBUGS and its R/S-PLUS interface BRugs. `http://www.stats.ox.ac.uk/pub/RWin/src/contrib`.

López, F. O. (2013). A Bayesian approach to parameter estimation in simplex regression model: a comparison with beta regression. *Revista Colombiana de Estadística 36*(1), 1–21.

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

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

Qiu, Z., P. X.-K. Song, and M. Tan (2008). Simplex Mixed-Effects Models for Longitudinal Proportional Data. *Scandinavian Journal of Statistics 35*(4), 577–596.

Raftery, A., M. Newton, J. Satagopan, and P. Krivitsky (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion). In J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (Eds.), *Bayesian Statistics 8*, Volume 8, pp. 1–45. Oxford University Press.

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

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

Song, P. X.-K., Z. Qi, and M. Tan (2004). Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal 46*(5), 540–553.

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

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

Thomas, A., B. O'Hara, U. Ligges, and S. Sturtz (2006). Making BUGS open. *R News 6*(1), 12–17.

Verkuilen, J. and M. Smithson (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 58*(4), 739–750.

Xie, F.-C., J.-G. Lin, and B.-C. Wei (2014). Bayesian zero-inflated generalized poisson regression model: estimation and case influence diagnostics. *Journal of Applied Statistics 41*(6), 1383–1392.

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

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

Appendix A

BUGS code to implement the ZOAB-Re model with covariates in p_0 and p_1

```
model
{
Cte<-1000000

for(i in 1:n)
{
b[i]~dnorm(0,tau)
}

for(j in 1:N)
  {
   zeros[j]        <- 0
   zeros[j]        ~  dpois(zeros.means[j])
   zeros.means[j]  <- -lBetaInf[j]+Cte
   fdBeta[j]       <- exp(loggam(a1[j]+a2[j])-loggam(a1[j])-loggam(a2[j])
                      +(a1[j]-1)*log(Y[j])+(a2[j]-1)*log(1-Y[j]))
   a1[j]           <- mu[j]*phi
   a2[j]           <- (1-mu[j])*phi
   logit(mu1[j])   <- Beta[1]  + Beta[2] * gender[j] + Beta[3]  * age[j]
                      + Beta[4] * hba1cd[j] + Beta[5] * smoker[j]
                      + Beta[6]  * incisor [j] + Beta[7] * premolar[j]
                      + Beta[8] * molar[j] + b[cluster[j]]
   mu[j]           <- max(0.00001,min(0.9999,mu1[j]))
   logit(p0[j])    <- gamma[1]  + gamma[2] * gender[j] + gamma[3] * age[j]
                      + gamma[4] * hba1cd[j] + gamma[5] * smoker[j]
                      + gamma[6] * incisor [j] + gamma[7] * premolar[j]
                      + gamma[8] * molar[j]
   logit(p1[j])    <- rho[1]    + rho[2]   * gender[j] + rho[3]   * age[j]
```

```
                      + rho[4] * hba1cd[j] + rho[5] * smoker[j]
                      + rho[6] * incisor [j] + rho[7] * premolar[j]
                      + rho[8] * molar[j]
   e[j]            <- equals(Y[j],0.0001) #zeros observados
   d[j]            <- equals(Y[j],0.9999) #uns observados
   fdBetaInf1[j]   <- (e[j]*p0[j]+d[j]*p1[j]
                      +(1-e[j])*(1-d[j])*fdBeta[j]*(1-p0[j]-p1[j]))
                      *step(1-p0[j]-p1[j])
   fdBetaInf[j]    <- max(0.00000001,fdBetaInf1[j])
   lBetaInf[j]     <- log(fdBetaInf[j])
   }

#Prioris para os parâmetros do modelo
for (i in 1:8)
  {
   Beta[i]  ~ dnorm(0,0.001)
   gamma[i] ~ dnorm(0,0.001)
   rho[i]   ~ dnorm(0,0.001)
  }

phi      ~  dgamma(0.1,0.01)
tau      <- pow(sigmabsd,-2)
sigmabsd ~  dunif(0,100)
sigma2b  <- pow(sigmabsd,2)
}
```

Appendix B

Simulation results obtained when the sample is generated from ZOAB-RE model

In this case, it was conducted a finite sample simulation study to investigate the consequences on parameter estimates after applying the LS transformation to the data in $[0,1]$. For the data generation scheme, there were generated 100 samples of the ZOAB-RE model with 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 multinomial draws.

	ZOAB-RE model				LS Beta			
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	0,27	0,15	0,14	0,07	0,15	0,14	0,15	0,12
β_1	0,35	0,22	0,16	0,11	0,17	0,17	0,15	0,13
ϕ	0,09	0,03	0,02	0,01	1,52	1,69	1,76	1,82
p_0	0,001	0,0001	0,0001	0,00001	-	-	-	-
p_1	0,001	0,0001	0,0003	0,0003	-	-	-	-
σ_b^2	0,96	0,97	0,98	1,05	12,81	12,87	12,91	12,95
CP								
β_0	0,98	0,93	0,97	0,98	0,82	0,51	0,53	0,28
β_1	0,97	0,94	0,95	0,96	0,87	0,59	0,61	0,38
ϕ	0,92	0,95	0,96	0,95	0,00	0,00	0,00	0,00
p_0	0,94	0,96	0,98	0,98	-	-	-	-
p_1	0,81	0,76	0,72	0,67	-	-	-	-
σ_b^2	0,90	0,60	0,49	0,31	0,00	0,00	0,00	0,00

Table B.1: Relative bias, MSE and CP for the parameters of ZOAB-RE and LS beta models using different sample size .

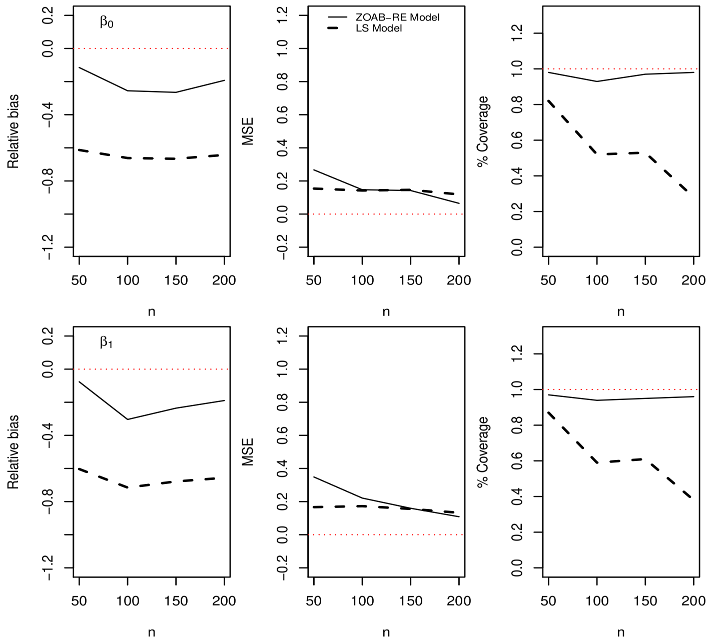

Figure B.1: Relative bias, MSE and CP for the parameters in ZOAB-RE and LS beta models.

Appendix C

Sensibility analysis for the hiperparameter of Dirichlet distribution

In order to analyse the sensibility of the hiperparameter of Dirichlet distribution, It was generated 50 sample of normal logistic distribution augmented by zeros and ones with parameters values $\beta_0 = -0.5$, $\beta_1 = 0.5$, $p_0 = 0.1$, $p_1 = 0.08$ and $\sigma_b^2 = 2$. Tables C.1 and C.2 present the results about the relative bias and MSE when it is used the prior Gamma$(1, 0.01)$ and Gamma$(0.1, 0.01)$, respectively, in the models ZOAS-RE and ZOAB-RE.

	ZOAS-RE model							
	Rel. bias				MSE			
	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
β_0	-0.10	-0.03	-0.08	-0.11	0.02	0.03	0.02	0.02
β_1	-0.09	-0.11	-0.10	-0.11	0.06	0.03	0.03	0.02
p_0	0.05	0.03	0.01	-0.0009	0.0003	0.0002	0.0001	0.00007
p_1	0.04	0.01	0.02	0.01	0.0002	0.0001	0.00009	0.00007
σ_b^2	-0.17	-0.22	-0.23	-0.23	0.32	0.26	0.26	0.25
	ZOAB-RE model							
	Rel. bias				MSE			
	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
β_0	-0.23	-0.16	-0.20	-0.23	0.05	0.03	0.02	0.02
β_1	-0.20	-0.24	-0.22	-0.22	0.06	0.04	0.02	0.02
p_0	0.05	0.03	0.01	-0.0009	0.0003	0.0002	0.0001	0.00007
p_1	0.04	0.01	0.02	0.01	0.0002	0.0001	0.00009	0.00007
σ_b^2	-0.36	-0.40	-0.41	-0.41	0.60	0.68	0.70	0.69

Table C.1: Relative bias and MSE of the parameters in the ZOAS-RE and ZOAB-RE models obtained with the prior $\alpha \sim Gamma(1, 0.01)$ in the hiperparameter of Dirichlet distribution

Modelo ZOAS-RE

	Rel. bias				MSE			
	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
β_0	-0.02	-0.06	-0.08	-0.10	0.05	0.02	0.02	0.02
β_1	-0.09	-0.06	-0.06	-0.16	0.05	0.03	0.02	0.03
p_0	0.03	-0.02	-0.01	-0.01	0.0003	0.0001	0.0001	0.00007
p_1	0.03	0.04	-0.03	-0.01	0.0002	0.0002	0.0001	0.00007
σ_b^2	-0.18	-0.21	-0.22	-0.24	0.32	0.23	0.26	0.25

Modelo ZOAB-RE

	Rel. bias				MSE			
	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
β_0	-0.15	-0.19	-0.21	-0.22	0.04	0.03	0.02	0.02
β_1	-0.20	-0.20	-0.20	-0.26	0.04	0.03	0.02	0.03
p_0	0.03	-0.03	-0.01	-0.01	0.0003	0.0002	0.0001	0.00007
p_1	0.03	0.04	-0.03	-0.01	0.0002	0.0002	0.0001	0.00007
σ_b^2	-0.37	-0.39	-0.41	-0.42	0.62	0.64	0.69	0.72

Table C.2: Relative bias and MSE of the parameters in the ZOAS-RE and ZOAB-RE models obtained with the prior $\alpha \sim Gamma(0.1, 0.01)$ in the hiperparameter of Dirichlet distribution.

Appendix D

Results about of simulation scheme 1 in the Chapter 3.

ZOAS-RE model modelling p_0 e p_1 with covariates												
	Rel. bias				MSE				CP			
	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
β_0	-0.10	-0.10	-0.12	-0.10	0.06	0.03	0.02	0.02	0.94	0.94	0.92	0.95
β_1	-0.10	-0.11	-0.13	-0.08	0.08	0.04	0.04	0.02	0.94	0.94	0.90	0.91
σ_b^2	-0.18	-0.23	-0.22	-0.20	0.31	0.28	0.24	0.21	0.82	0.60	0.52	0.43
ZOAS-RE model with p_0 and p_1 constants across all observations												
	Rel. bias				MSE				CP			
	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
β_0	-0.10	-0.11	-0.12	-0.11	0.06	0.03	0.02	0.02	0.96	0.94	0.92	0.94
β_1	-0.10	-0.11	-0.13	-0.08	0.08	0.04	0.04	0.02	0.95	0.95	0.90	0.90
σ_b^2	-0.18	-0.23	-0.22	-0.20	0.31	0.28	0.24	0.21	0.82	0.60	0.52	0.43

Table D.1: Results of the scheme of simulation 1 presented in the chapter 3 where the data is analyzed of the ZOAS-RE model

ZOAB-RE model modelling p_0 and p_1 with covariates												
	Rel. bias				MSE				CP			
	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
β_0	-0.23	-0.22	-0.23	-0.22	0.06	0.03	0.03	0.02	0.90	0.88	0.81	0.80
β_1	-0.25	-0.23	-0.25	-0.19	0.07	0.04	0.04	0.02	0.92	0.90	0.86	0.85
σ_b^2	-0.36	-0.40	-0.40	-0.37	0.61	0.68	0.68	0.60	0.52	0.12	0.02	0.12
ZOAB-RE model with p_0 and p_1 constants across all observations												
	Rel. bias				EQM				CP			
	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200	n=50	n=100	n=150	n=200
β_0	-0.23	-0.22	-0.24	-0.22	0.06	0.03	0.03	0.02	0.91	0.89	0.81	0.80
β_1	-0.25	-0.23	-0.25	-0.19	0.07	0.04	0.04	0.02	0.93	0.89	0.86	0.84
σ_b^2	-0.36	-0.40	-0.40	-0.37	0.61	0.68	0.68	0.60	0.50	0.12	0.02	0.12

Table D.2: Results of the scheme of simulation 1 presented in the chapter 3 where the data analized by the ZOAS-RE model