Beta regression with a small cluster at a boundary

Daniel Araújo Nóbrega

Thesis presented to the
Institute of Mathematics and Statistics
of the University of São Paulo
in partial fulfillment
of the requirements
for the degree of
Master of Science

Program: Statistics
Advisor: Prof. Dr. Silvia Lopes de Paula Ferrari

During this work, the author was supported by CNPq

São Paulo
30th June, 2021

Beta regression with a small cluster at a boundary

Daniel Araújo Nóbrega

This is the original version of the
thesis prepared by the candidate
Daniel Araújo Nóbrega, as submitted
to the Examining Committee.

Acknowledgements

Maturity is the capacity to endure uncertainty.

— John Finley

First of all, I would like to thank my supervisor Silvia for the collaboration and advice throughout the process of writing this thesis an also thank the committee for the suggestions. I would also like to thank my parents for the support during my masters and all my colleagues during this time. Finally, I would like to thank CNPq for the financial support that helped make this all possible.

Resumo

Daniel Araújo Nóbrega. **Regressão beta com um pequeno cluster em uma fronteira**. Dissertação (Mestrado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2021.

Análises de dados cujas variáveis respostas estão contidas no intervalo (0,1) têm recebido muita atenção nas últimas duas décadas, principalmente com o uso do modelo de regressão beta. No entanto, existem situações em que os dados contêm observações nas fronteiras, isto, é observações iguais a zero ou a um, em que outras metodolgias precisam ser consideradas, Neste trabalho, o foco é em dados que têm um pequeno cluster de observações em uma das fronteiras e os métodos utilizados fornecem maneiras de continuar ajustando um modelo de regressão beta, por máxima verossimilhança ou por um método de estimação robusto, para estes cenários após uma adaptação dos dados ou usar um modelo que é capaz lidar com a presença de observações nas fronteiras; aqui, o modelo de regressão beta inflacionado e um modelo de quasi-verossimilhança foram usados para esta finalidade. Os métodos foram aplicados em dois conjuntos de dados com características distintas; análises de diagnóstico foram conduzidas para avaliar a qualidade dos ajustes e então, cenários de simulação foram feitos para avaliar a performance de cada um dos métodos em situações que podem surgir na prática. Finalmente, algumas conclusões foram apresentadas sobre quais métodos funcionam melhor em cada uma das situações exploradas.

Palavras-chave: Estimação robusta. Observações de fronteira. Quasi-verossimilhança. Regressão beta. Regressão beta inflacionada.

Abstract

Daniel Araújo Nóbrega. **Beta regression with a small cluster at a boundary**. Thesis (Masters). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2021.

Analyses of data that have response variables contained in the (0,1) interval have received a lot of attention in the past two decades, most notably through the use of the beta regression model. However, there are situtations where there are boundary observations in the data, i.e. observations equal to zero or to one, in which other methodologies must be considered. In this work, the focus is on data that have a small custer of observations at one of the boundaries and the methods used either provide ways to still fit a beta regression model, via maximum likelihood or via a robust estimation method, for these scenarios by adapting the data to fit onto the (0,1) interval or using a model that can naturally cope with the presence boundary observations; here, the inflated beta regression model and a quasi-likelihood model were used for this purpose. The methods were applied to two different datasets that had distinct characteristics; diagnostic analyses were conducted to assess the quality of the fits and then simulation scenarios were carried out to evaluate the performance of each of the methods in situations that may arise in practice. Finally, some conclusions were made about which methods work best in each of the situations explored.

Keywords: Beta regression. Boundary observations. Inflated beta regression. Quasi-likelihood. Robust estimation.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The analysis of data observed in the (0,1) interval has received plenty of attention in the past two decades through research that aims to develop methodologies specific to such data. Due to its bounded nature and different shapes it can assume, this type of data presents challenges that must be addressed by whichever method is chosen to analyse it.

Regression models have been proposed based on a number of distributions: the beta regression model proposed in Ferrari and Cribari-Neto (2004), the simplex model (Jørgensen, 1997), the Kumarasawamy regression model (Mitnik and Baek, 2013), among others, that are defined on the (0,1) interval, therefore these models provide a way to link a response variable to a set of explanatory variables that will always respect the bounded nature of the data and remain contained in (0,1). The beta regression proposed by Ferrari and Cribari-Neto (2004) is the most used regression model for the type of data in question as the beta distribution is very versatile and can be fitted to different shapes of data. Smithson and Verkuilen (2006) extend the beta regression model to accommodate the situation of having a varying precision. However, if the data in the unit interval also includes observations at the boundaries, i.e. observations equal to zero or equal to one, the aforementioned assumption that the density is greater than zero only in (0,1) is violated, thus these models are not, on their own, adequate to analyse this situation. That being the case, it is important to understand in what scenarios boundary observations may appear and what they can mean before proceeding to methods that allow these observations to be taken into account. For instance, the small cluster of observations at one of the boundaries may not have a special meaning at all, such as seen in Rudolf *et al.* (2019).

One possible scenario where there may be boundary observations in the data is when these imply a special meaning. For instance, suppose the response variable in analysis represents the proportion of people in a country that have been diagnosed with poliomyelitis in the past ten years. If an observation of this response is zero, it suggests that this disease may in fact have been eradicated in that particular country. However, if this observation is greater than zero, it can possibly indicate that the inhabitants of that country may still be susceptible to contracting that disease, especially if no eradication campaigns have been conducted since the diagnosed cases. In such situations, due to the special meaning of the boundary observations, a regression model based on a distribution that is defined on (0,1)

but is inflated to include at least one of the boundaries is often used. An example of such a model is the inflated beta regression model proposed by Ospina and Ferrari (2012), which provides a model that allows the fitting of a model where the response variable is contained in [0,1), (0,1] or even [0,1]. Another situation where boundary observations may arise is when these do not carry a special meaning. Rudolf *et al.* (2019) analyse the fracture resistance of dragonfly wings, where the measurement of fracture is given as a proportion of the wing that presents some kind of injury. In this dataset, there are a number of observations equal to zero, however given that the measurement instrument is mostly visual and subject to slight error, there is probably very little difference between a wing that is considered to be without any injuries (response equal to zero) and a wing considered 0.1% compromised. Thus, making it a different scenario to when there is a particular meaning to the boundary observations, so it would be natural to hypothesize that the methods to analyse these different scenarios may differ.

Another aspect to consider is the quantity of boundary observations compared to the total sample size in the data. If an observation being equal to zero (or to one) is a regular occurrence and it represents a significant proportion of the data, it makes sense to analyse these observations separately as these observations may have different characteristics from the rest of the data and having a large number of them allows statistical methods to be used specifically on these observations, such as fitting the aforementioned inflated regression models. As a regular occurrence of boundary observations implies a clear way of analysing the data, the focus of this work will be on situations in which the amount of boundary observations is relatively small compared to the full data as in such cases the best method to use will be more subjective and may vary depending on the characteristics of the dataset.

The aim of this work will be to compare different methodologies that can deal with having small clusters of observations at the boundaries and offer conclusions taking into account the context of the data in which these methods will be applied to. Naturally, different methods may work the best depending on the context of the data and how the boundary observations compare to the rest of the data. It is important to note that there are other ways to analyse data with boundary observations by transforming the response variable in order for it to be normally distributed, such as in Piepho (2003) and Malik and Piepho (2016), but the focus here will be on methodologies that allow a model related to the beta regression model to be fitted.

In Chapter 2, the methodologies that will be compared are presented, along with the theory behind some of them and the benefits and drawbacks of choosing each one. The methodologies are applied to two distinct datasets in Chapter 3, in which this distinction affects what methods may be appropriate for use in each one. All computational work was done using the R software (R Core Team, 2020) and the source code for the applications can be found at https://github.com/danielanobrega/BR_boundaries. In Chapter 4, three simulation scenarios are conducted to further assess the performance of the methods and to expand the findings in the applications chapter. Finally, the conclusions are presented in Chapter 5.

Chapter 2

Methods

Data observed in the [0,1] interval may or may not have a negligible number of observations equal to zero or one, therefore it is important to consider different approaches to analyse this type of data.

There are situations in which the number of boundary observations is either sufficiently negligible to not consider it a mass of points large enough to warrant a probability distribution associated with them or there may be no particular special meaning relative to the appearance of these observations in the sample.

A common way to deal with having zeros or ones in the data is to transform either all the observations or just those at the boundary in order for all of them to be contained within the (0, 1) interval. The advantage of doing this is that it allows methodologies developed for analysis of data in the open interval (0,1) to be used even in the presence of boundary observations, most notably, the beta regression model introduced by Ferrari and Cribari-Neto (2004) and its extensions because of the versatility of the beta distribution, which allows it to model data with different shapes.

Let $y_1, \ldots, y_n$ be a random sample, in which each y_i, $i = 1, \ldots, n$, follows a beta distribution with mean μ_i ($0 < \mu_i < 1$) and precision parameter $\phi_i > 0$ (we write $y_i \sim \mathcal{B}(\mu_i, \phi_i)$). The probability density function of y_i is

$$f(y_i; \mu_i, \phi_i) = \frac{\Gamma(\phi_i)}{\Gamma(\mu_i\phi_i)\Gamma((1-\mu_i)\phi_i)} y^{\mu_i\phi_i - 1}(1-y_i)^{(1-\mu_i)\phi_i - 1}, \quad y_i \in (0,1), \tag{2.1}$$

where $\Gamma(\cdot)$ is the gamma function. If $y_i \sim \mathcal{B}(\mu_i, \phi_i)$, then $\text{Var}(y_i) = V(\mu_i)/(\phi_i + 1)$, where $V(\mu_i) = \mu_i(1-\mu_i)$ denotes the "variance function". For a fixed value of μ_i, the larger the value of the precision parameter ϕ_i, the smaller the variance of y_i.

The beta regression (BR) model is defined by assuming that the mean of y_i and the precision are

$$g_\mu(\mu_i) = \eta_{1i} = x_i^\top \beta \tag{2.2}$$

and

$$g_\phi(\phi_i) = \eta_{2i} = z_i^\top \gamma \tag{2.3}$$

respectively, where $g_\mu(\cdot)$ and $g_\phi(\cdot)$ are strictly monotonic and twice differentiable link functions that map η_{1i} and η_{2i} onto the (0, 1) and (0, ∞) intervals, respectively, $x_i^\top = (x_{i1}, \ldots, x_{iq_1})$ and $z_i^\top = (z_{i1}, \ldots, z_{iq_2})$ are vectors of covariates associated with the ith observation in the sample and $\beta = (\beta_1, \ldots, \beta_{q_1})^\top \in \mathbb{R}^{q_1}$, and $\gamma = (\gamma_1, \ldots, \gamma_{q_2})^\top \in \mathbb{R}^{q_2}$ are vectors of unknown regression parameters. The model proposed in Ferrari and Cribari-Neto (2004) is a particular case of the described model in which $\phi_i = \phi \ \forall i = 1, \ldots, n$, that is say that is a constant precision model, which may in fact be adequate in certain situations.

The estimation of the parameters in the beta regression models can be done by the maximum likelihood estimator, therefore the log-likelihood function $\ell(\beta, \gamma) = \sum_{i=1}^n \ell_i(\mu_i, \phi_i)$ must be maximized, where

$$\ell_i(\mu_i, \phi_i) = \log\Gamma(\phi_i) - \log\Gamma(\mu_i\phi_i) - \log\Gamma((1-\mu_i)\phi_i) + (\mu_i\phi_i - 1)\log y_i + \{(1-\mu_i)\phi_i - 1\}\log(1-y_i).$$

There are situations, however, where the maximum likelihood estimator is very sensitive to the presence of discrepant observations in the data. Ribeiro and Ferrari (2020) proposed a robust estimation method for beta regression models based on the maximization of a reparameterized L_q-likelihood. This alternative estimator offers a trade-off between robustness and efficiency through a tuning constant q, whose optimal value is selected by using a data driven method that ensures full efficiency of the estimator in the absence of outliers. Henceforth, this method provides a robust beta regression inference.

2.1 Mapping data onto (0,1)

As previously mentioned, before using the beta regression model on data that possesses boundary observations, it is necessary to use a method to map the response variable onto the (0,1) interval. The methods are more adequate to use in situations where there is not a large mass of observations equal to zero or to one and when there is no special meaning to the boundary observations that differentiates them from responses with values close to those of the boundaries. For instance, let the response value be the percentage of votes a certain candidate had in a particular district. There is not much difference between having 0% of the votes and 0.1%, as both cases would just imply that the candidate is very unpopular in that particular district.

The subsequent subsections will cover some of the most commonly used methods to transform data into the (0,1) interval, consequently presenting ways to make it possible to model the data with a beta regression model which is the most popular when it comes to dealing with responses in the unit interval.

2.1.1 Removing boundary observations

A common method to transform the data, albeit a little extreme, is to simply remove the boundary observations, thus only keeping sample units in the (0,1) interval and

then proceeding with the desired analysis methodology. Although this method may be considered simple and fairly intuitive in situations where there are but few boundary observations, discarding observations may not be the most appropriate solution as it can be valuable to a better understanding of the situation being studied.

2.1.2 Adding or subtracting $\varepsilon > 0$

A very simple way to transform data in the [0,1] interval so that they fit in the (0,1) interval is to add (subtract) a very small value $\varepsilon > 0$ to observations equal to zero (to one), thereby altering the data so that it no longer has any boundary observations, furthermore allowing the use of methodologies to analyse data in the (0,1) interval.

In Rudolf *et al.* (2019) this method was used; in order to fit the beta regression model, the authors opted to add 0.0001 to the observations equal to zero in the dataset. This method is also used to deal with boundary observations in Hunger *et al.* (2012) and in Lima-Filho *et al.* (2020). In Ribeiro and Ferrari (2020), 0.001 was subtracted from the sole observation equal to one before proceeding with the regression model. The same dataset will be analysed here in Chapter 3.

It is important to keep in mind that how small the ε must be depends on the magnitude of the response variable's observations, thus choosing an inappropriate value of ε may negatively affect not only the fitting of the model, but also the interpretations.

2.1.3 Replacing by the largest (smallest) observations in the unit interval

It is also possible to replace the boundary observations to maintain the range of the observations in the (0,1) interval.

Let the sample to be analysed be given by $\mathcal{Y} = \{y_1, \ldots, y_n\}$, where n is the sample size and the observations are contained within the $[0, 1]$ interval.

Let $y_1^{(0)}, \ldots, y_k^{(0)}$ be observations from the sample that are equal to zero and $y_1^{(1)}, \ldots, y_{k'}^{(1)}$ observations from the sample equal to one.

To transform the data, we modify the observations at the boundaries. For $j = 1, \ldots, k$ and $j' = 1, \ldots, k'$ let

$$y_j^{(0)} = \min_i \{ y_i \in \mathcal{Y} \, : \, 0 < y_i < 1 \}$$

and

$$y_{j'}^{(1)} = \max_i \{ y_i \in \mathcal{Y} \, : \, 0 < y_i < 1 \}.$$

That way all observations will be in the unit interval and further analyses may proceed by using a method appropriate for data in the (0,1) interval.

2.1.4 Linear transformation

Smithson and Verkuilen (2006) analysed the relative contribution of nonverbal IQ and dyslexic versus non-dyslexic status to the distribution of 44 children's (25 dyslexic and 19 non-dyslexic) scores on a test of reading accuracy. Verifying the skewness of the data at hand, they decided to linearly transform the scores from the test from their original scale to the open interval (0,1) and avoid 0's and 1's by taking

$$y' = (1 - 1/n)y + 1/2n \tag{2.4}$$

where y denotes the original score and n is the sample size. Note that in this transformation, all observations in the samples are transformed, unlike the previously mentioned methodologies to avoid boundary observations. This transformation works as the weighted mean between the observation y and 1/2, which is the centre of the (0,1) interval, with weight (1-1/n) and 1/n, respectively.

Morrison *et al.* (2020) analysed conscientious vaccination exemptions in kindergarten in the US state of Texas and this data contained boundary observations, therefore before modelling the data, this transformation was used. In fact, this transformation seems to be the most common way in published works to transform data onto the (0,1) interval before fitting a model; this transformation is also discussed as a way of dealing with boundary observations in Schmid *et al.* (2013) and in Hunger *et al.* (2012).

Thus, it is possible to deal with data in the unit interval that contain observations equal to 0 or 1 by transforming the data using the method proposed by Smithson and Verkuilen (2006), which, in this paper, will hereafter be referred to as the linear transformation.

2.2 Models that accommodate boundary observations

Another way to deal with boundary observations in the data, is to use models that can naturally cope with their presence, thus not requiring a transformation of the data before proceeding with the fitting of a model.

2.2.1 Zero-inflated beta regression and one-inflated beta regression

In situations in which the number of boundary observations is non-negligible, it may be worth investigating if the boundary observations have any special meaning which separates them from the rest of the data. For instance, suppose the variable being analysed refers to the proportion of monthly income spent with mobile phone services. As this is a proportion, the value of each observation will be between 0 and 1. It is impossible for an observation to be equal to 1 seeing as, at the very least, part of the monthly income will be spent with food, bills and other necessities. However, it is possible that there may be

observations equal to zero in the data, furthermore these have a special meaning, which is that the sample unit (whether it be a household or an individual) does not have a mobile phone, hence separating them from the rest of the data. Because of this distinct nature between the boundary observations and the remaining ones, such as in the aforementioned scenario, it might be worth analysing this type of data separating the interior and boundary observations by resorting to two-part models.

By definition, two-part models are able to separate the analysis of two different parts of the data, therefore allowing the interpretation of the results for the interior and boundary observations to be conducted separately. In the scenario previously described, a two-part model would grant the possibility of interpreting the results for those who do not have a mobile phone to be done independently from those who do. In this work, the two-part model used will be one that is defined under the assumption that the interior observations are distributed according to the beta distribution. To accommodate the boundary observations, the two-part model used will be based on the inflated beta distribution initially proposed by Ospina and Ferrari (2010).

The inflated beta distributions are mixed continuous-discrete distributions. The continuous component has a beta distribution and the discrete component is a degenerate component at zero or at one, called the zero-inflated beta distribution and one-inflated beta distribution, respectively, or a Bernoulli distribution for the discrete component (if both zeros and ones are present in the sample) mixed with a beta distribution for the continuous component, which is then named the zero-and-one inflated beta distribution.

The cumulative distribution function of the beta distrbution inflated at c, where $c = 0$ or $c = 1$, is given by

$$\mathrm{BI}_c(y; \alpha, \mu, \phi) = \alpha \mathbb{I}_{[c,+\infty]}(y) + (1 - \alpha) F(y; \mu, \phi), \;\; y \in \mathbb{R}, \tag{2.5}$$

where $\mathbb{I}_A(y)$ is an indicator function that equals 1 if $y \in A$ and 0 if $y \notin A$ and $F(\cdot)$ denotes the cumulative distribution of the beta distribution $\mathcal{B}(\mu, \phi)$. Also, $0 < \alpha < 1$ is the mixture parameter. The corresponding probability density function with respect to the measure generated by the mixture is given by

$$\mathrm{bi}_c(y; \alpha, \mu, \phi) = \begin{cases} \alpha, & \text{if } y = c, \\ (1 - \alpha) f(y; \mu, \phi), & \text{if } y \in (0, 1) \end{cases} \tag{2.6}$$

where $f(y; \mu, \phi)$ denotes the density of the beta distribution $\mathcal{B}(\mu, \phi)$ in (2.1).

If $c = 0$, the distribution is called zero-inflated beta distribution (BEZI) and we write $y \sim \mathrm{BEZI}(\alpha, \mu, \phi)$. In the case that $c = 1$, the distribution is called one-inflated beta distribution (BEOI) and we write $y \sim \mathrm{BEOI}(\alpha, \mu, \phi)$. Some authors refer to these distributions as the zero-augmented beta distribution and the one-augmented beta distribution, respectively, such as in Douma and Weedon (2019).

The rth moment of y and its variance are

$$\mathrm{E}(y^r) = \alpha c + (1 - \alpha)\mu_r, \;\; r = 1, 2, \ldots,$$

and

$$\mathrm{Var}(y) = (1-\alpha)\frac{V(\mu)}{\phi+1} + \alpha(1-\alpha)(c-\mu)^2,$$

respectively, where $\mu_r = (\mu\phi)_{(r)}/\phi_{(r)}$, with $a_{(r)} = a(a+1)\ldots(a+r-1)$, is the rth moment of the beta distribution. The mean of the distribution is a weighted average between the degenerate distribution at c and the beta distribution.

Let $y_1, \ldots, y_n$ be a random sample, where each y_i follows a c-inflated beta distribution with parameters α_i, β_i and ϕ_i. The likelihood function for $\theta = (\alpha, \mu, \phi)$, where $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$, $\mu = (\mu_1, \ldots, \mu_n)^\top$ and $\phi = (\phi_1, \ldots, \phi_n)$ given the sample $(y_1, \ldots, y_n)$ is

$$L(\theta) = \prod_{i=1}^{n} \mathrm{bi}_c(y_i; \alpha_i, \mu_i, \phi_i) = L_1(\alpha)L_2(\mu, \phi),$$

where

$$L_1(\alpha) = \prod_{i=1}^{n} \alpha_i^{\mathbb{I}_{\{c\}}(y_i)}(1-\alpha_i)^{1-\mathbb{I}_{\{c\}}(y_i)},$$

and

$$L_2(\mu, \phi) = \prod_{i=1}^{n} f(y_i; \mu_i, \phi_i)^{1-\mathbb{I}_{\{c\}}(y_i)}.$$

The likelihood function is said to be separable as it factors into a term that depends only on α_i and one that depends only on μ_i and ϕ_i. Therefore, the maximum likelihood estimation for (μ_i, ϕ_i) can be performed separately to that of α_i as if the other value was known.

The log-likelihood function for the inflated beta distribution is given by

$$\ell(\theta) = \log L(\theta) = \ell_1(\alpha) + \ell_2(\mu, \phi),$$

where

$$\ell_1(\alpha) = \sum_{i=1}^{n} \log\left\{\alpha_i^{\mathbb{I}_{\{c\}}(y_i)}(1-\alpha_i)^{1-\mathbb{I}_{\{c\}}(y_i)}\right\}$$

and

$$\ell_2(\mu, \phi) = \sum_{i=1}^{n}\{1-\mathbb{I}_{\{c\}}(y_i)\}\log f(y_i; \mu_i, \phi_i).$$

Let $y_1, \ldots, y_n$ be independent random variables in which y_i follows, $\forall i = 1, \ldots, n$, either the zero-inflated beta distribution or the one-inflated beta distribution with conditional mean $E(y_i|y_i \in (0,1)) \equiv \mu_i$ and unknown mixture parameter α. The regression model is defined under the assumptions (2.2) and (2.3). It is also assumed that

$$g_\alpha(\alpha_i) = w_i^\top \alpha,$$

where $w_i^\top = (w_{i1}, \dots, w_{iq_3})$. However, in the applications on this thesis α was assumed to be constant throughout the observations as in the situations analysed here there are too few observations at a boundary to justify a submodel for α. Although it could be argued that the assumption that α is constant is unreasonable, since there are but few observations equal to zero, this assumption is not likely to cause problems in the model. The parameters can be estimated by maximizing $L(\theta)$ (or equivalently, $\ell(\theta)$). The model proposed in Ospina and Ferrari (2012) is a particular case of the described model in which $\phi_i = \phi \ \ \forall i = 1, \dots, n$.

This model is implemented in the `gamlss` package on `R` and even though the estimation can be made using the maximum likelihood estimator, the package uses its own algorithm to estimate parameters, thus estimations can differ very slightly.

The method proposed in this subsection is not suitable for the analysis of data that possess observations equal to zero and equal to one. Hence an alternative method is needed for such situations. Ospina and Ferrari (2010) also introduce the zero-and-one inflated beta distribution, which is appropriate for such scenarios, however it will not be expanded upon here since we are not dealing with such scenarios.

2.2.2 Quasi-beta regression model

To deal with the presence of boundary observations one might also resort to the quasi-likelihood (QL) approach. Due to the fact that a distribution for the data is not assumed when this methodology is employed, the presence of observations equal to zero or to one does not violate any prior assumptions, which would be the case if a distribution defined in the (0, 1) interval was chosen to model the data.

There are various regression models for data in the (0,1) interval, such as the, already mentioned, beta regression model and the simplex regression model discussed in Kieschnick and Mccullough (2003). There have also been proposed new probability density functions for modelling continuous bounded data. Lemonte and Bazán (2016) introduced a model based on the Johnson S_B distribution and, as an alternative to the beta regression model, Mitnik and Baek (2013) proposed a regression model based on the Kumaraswamy distribution. There have also been models that have been proposed using mixtures of beta distributions that alow more flexibility in dealing with atypical observations, such as in Bayes *et al.* (2012), Migliorati *et al.* (2017) and Di Brisco *et al.* (2020). Therefore, when analysing data in the unit interval, there is a certain variety of possible models to choose from. Oftentimes, however, it may be difficult to argue which is the best model for a particular dataset with conviction seeing as different measures of goodness-of-fit (Akaike Information Criterium and Bayesian Information Criterium, for instance) may lead to different models being considered the "best" one. In cases where the choice may prove to be unclear, it may be advantageous to possess a regression model that can adapt to different forms of the variance function, therefore being flexible to different shapes of the response variable's distribution and that can be easily implemented in practice.

Bonat *et al.* (2019) propose a quasi-beta regression model based only on second moment assumptions. Let $y_1, \dots, y_n$ be independent random variables in which the distribution of

y_i need not be specified. The model is defined under the assumptions that the mean is as defined in (2.2) and that

$$\text{Var}(y_i) = \sigma_i = \sigma\mu_i^p(1-\mu_i)^p,$$

where σ is a dispersion parameter and p is a power parameter that allows more flexibility in modelling the relationship between the mean and the variance function. For $p = 1$ the variance is equivalent to that of the beta distribution, where $\sigma = 1/(1+\phi)$. For convenience, the vector of parameters used in this model will be denoted by $\theta = (\beta^\top, \lambda^\top)^\top$, where $\lambda^\top = (\phi, p)^\top$.

Adapting results presented in Jørgensen and Knudsen (2004) and Bonat and Jørgensen (2015), the authors adopted the quasi-score and Pearson estimating functions for estimation of the regression and dispersion parameters. The quasi-score function for β is

$$\psi_\beta(\beta,\lambda) = \left(\sum_{i=1}^n \frac{\partial\mu_i}{\partial\beta_1}\sigma_i^{-1}(y_i-\mu_i), \ldots, \sum_{i=1}^n \frac{\partial\mu_i}{\partial\beta_q}\sigma_i^{-1}(y_i-\mu_i)\right)^\top,$$

where $\partial\mu_i/\partial\beta_j = \mu_i(1-\mu_i)x_{ij}$ for $j = 1, \ldots, q$. The entry (j,k) of the $q \times q$ sensitivity matrix S_β for ψ_β is

$$\text{S}_{\beta_{jk}} = E\left(\frac{\partial}{\partial\beta_k}\psi_{\beta_j}(\beta,\lambda)\right) = -\sum_{i=1}^n \mu_i(1-\mu_i)x_{ij}\sigma_i^{-1}x_{ik}\mu_i(1-\mu_i)$$

and the entry (j,k) of the $q \times q$ variability matrix V_β for ψ_β is

$$\text{V}_{\beta_{jk}} = \text{Cov}(\psi_{\beta_j}(\beta,\lambda), \psi_{\beta_k}(\beta,\lambda)) = \sum_{i=1}^n \mu_i(1-\mu_i)x_{ij}\sigma_i^{-1}x_{ik}\mu_i(1-\mu_i).$$

The Pearson estimating functions for the dispersion parameters are

$$\psi_\lambda(\lambda,\beta) = \left(-\sum_{i=1}^n \frac{\partial\sigma_i^{-1}}{\partial\phi}[(y_i-\mu_i)^2-\sigma_i],\ -\sum_{i=1}^n \frac{\partial\sigma_i^{-1}}{\partial p}[(y_i-\mu_i)^2-\sigma_i]\right)^\top.$$

These are unbiased estimating functions for λ based on the square residuals $(y_i-\mu_i)^2$ with expected value σ_i.

The entry (j,k) of the 2×2 sensitivity matrix S_λ is

$$S_{\lambda_{jk}} = \text{E}\left(\frac{\partial}{\partial\lambda_k}\psi_{\lambda_j}(\lambda,\beta)\right) = -\sum_{i=1}^n \frac{\partial\sigma_i^{-1}}{\partial\lambda_j}\sigma_i\frac{\partial\sigma_i^{-1}}{\partial\lambda_k}\sigma_i,$$

where λ_j or λ_k denote either ϕ or p. The cross entries of the sensitivity matrices $S_{\beta\lambda}$ and $S_{\lambda\beta}$ are

$$S_{\beta_j\lambda_k} = \text{E}\left(\frac{\partial}{\partial\lambda_k}\psi_{\beta_j}(\beta,\lambda)\right) = 0$$

and

$$S_{\lambda_j\beta_k} = \mathrm{E}\left(\frac{\partial}{\partial\beta_k}\psi_{\lambda_j}(\lambda,\beta)\right) = -\sum_{i=1}^{n}\frac{\partial\sigma_i^{-1}}{\partial\lambda_j}\sigma_i\frac{\partial\sigma_i^{-1}}{\partial\beta_k}\sigma_i.$$

Therefore, the joint sensitivity matrix for θ is

$$S_\theta = \begin{pmatrix} S_\beta & 0 \\ S_{\lambda\beta} & S_\lambda \end{pmatrix}.$$

The asymptotic variance of the estimating function estimators $\widehat{\theta}$ is obtained through the inverse Godambe information matrix, whose general form is $J_\theta^{-1} = S_\theta^{-1}V_\theta S_\theta^{-\top}$, where $-\top$ indicates the inverse transpose operation and

$$V_\theta = \begin{pmatrix} V_\beta & V_{\beta\lambda} \\ V_{\lambda\beta} & V_\lambda \end{pmatrix},$$

where $V_{\lambda\beta} = V_{\beta\lambda}^{\top}$ and V_λ depend on the third and fourth moments of y_i, respectively. To avoid this dependance on higher moments, Bonat *et al.* (2019) use empirical versions of V_λ and $V_{\lambda\beta}$ whose entries (j,k) are given by

$$\widetilde{V}_{\lambda_{jk}} = \sum_{i=1}^{n}\psi_{\lambda_j}(\lambda,\beta)_i\psi_{\lambda_k}(\lambda,\beta)_i$$

and

$$\widetilde{V}_{\lambda_j\beta_k} = \sum_{i=1}^{n}\psi_{\lambda_j}(\lambda,\beta)_i\psi_{\beta_k}(\beta,\lambda)_i.$$

The approximate distribution of $\widehat{\theta}$ is the multivariate Gaussian distribution with mean θ and variance J_θ^{-1}. The task of estimating these functions have been implemented onto the `mcglm` package on `R` (R Core Team, 2020). Recently, during the writing of this thesis, the package was removed from the `CRAN` repository on `R`, but the results in this work are not affected and the package may return to `R`. An alternative to this model if you consider p fixed as one, which would be the classical quasi-likelihood approach for bounded data, is to use the quasibinomial family of the `glm` function on `R` as it provides estimates that are nearly equivalent to the method discussed in this section.

One limitation of this model is that it has been implemented solely as a constant precision (or dispersion) model, therefore in situations where it may be appropriate to fit a model with varying precision, this model is not adequate.

Chapter 3

Applications

In this chapter, two applications will be considered to illustrate the differences between the aforementioned methods to deal with having boundary observations for data in the unit interval. The first one showcases an example where the boundary observation present in the dataset is a discrepant observation, therefore the robust beta regression model introduced in Ribeiro and Ferrari (2020) is also fitted. The second application is an example where the boundary observations are very similar in value to the observations in (0,1), thus providing a different context for the data.

In all applications, the `R` software (R Core Team, 2020) was used to fit models and produce figures and tables. In order to fit the beta regression models, the `betareg` package was used, the robust beta regression models (RobBR) was fitted with computational programs provided in Ribeiro and Ferrari (2020) which can be found in https://github.com/terezinharibeiro/RobustBetaRegression. The inflated beta regression models are implemented in the `gamlss` package and the quasi-likelihood model was fitted using the `mcglm` package.

For the diagnostics of the models, worm plots were created. According to Buuren and Fredriks (2001), a worm plot is a diagnostic tool for checking the residuals within different ranges of the explanatory variables. In order to produce the worm plots the randomized quantile residual given by

$$r_{q,i} = \Phi^{-1}\{F(y_i; \hat{\mu}_i, \hat{\phi}_i)\}$$

was used, where $F(\cdot)$ is the cumulative distribution function of the beta regression defined in (2.1), $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution and $\hat{\mu}_i$ and $\hat{\phi}_i$ are the estimates for μ_i and ϕ_i respectively. In the case of the zero-inflated beta regression model the randomized quantile residual is given by

$$r_{q,i} = \begin{cases} \Phi^{-1}\{\hat{\alpha} u_i\}, & \text{if } y_i = 0, \\ \Phi^{-1}\{\hat{\alpha} + (1-\hat{\alpha}) F(y_i; \hat{\mu}_i, \hat{\phi}_i)\}, & \text{if } y_i \in (0,1), \end{cases} \tag{3.1}$$

where u_i is a random draw from the uniform distribution in the (0,1) interval. The ran-

domized quantile residual in the one-inflated beta regression model is

$$r_{q,i} = \begin{cases} \Phi^{-1}\{(1-\widehat{\alpha})F(y_i;\widehat{\mu}_i,\widehat{\phi}_i)\}, & \text{if } y_i \in (0,1), \\ \Phi^{-1}\{u_i\}, & \text{if } y_i = 1, \end{cases} \tag{3.2}$$

where u_i is a random draw from the uniform distribution in $(1-\widehat{\alpha}, 1)$. As per Dunn and Smyth (1997), apart from the sampling variability in $\widehat{\mu}_i$ and $\widehat{\phi}_i$, the $r_{q,i}$ is exactly standard normal, therefore if β and γ are consistently estimated, $r_{q,i}$ converges in distribution to a standard normal distribution. In Pereira (2019), it is stated that when using the beta regression, the randomized quantile regression is overall a better choice to perform diagnostics analysis, compared to the standardized weighted residual 1 and 2, particularly when the observations are very close to zero or very close to one.

There are also additional diagnostics graphs that can be found in Appendix A and in Appendix B. For the beta regression and robust beta regression models, the standardized weighted residual 2 proposed by Espinheira *et al.* (2008) was used to produce the diagnostic graphs aside from the worm plots. The expression for this residual can be found in eq. (7) of that paper. In order to perform the same diagnostic techniques in both applications, which have different characteristics, the randomized quantile residual and the standardized residual type 2 were both used.

The QL model estimates the value of a dispersion parameter $\sigma = 1/(1+\phi)$, where ϕ is the precision parameter of the beta distribution as presented in (2.1). To unify the estimates so that they may be compared, the dispersion parameter of the QL models was transformed to be on the same scale as the ϕ in the beta regression models (and the inflated beta regression models) by taking $(1-\widehat{\sigma})/\widehat{\sigma}$ and the standard error of the estimate of γ was obtained through the delta method.

3.1 Tuna application

In this application, the dataset used can be found in the Supplementary Material of Monllor-Hurtado *et al.* (2017). The response variable is the tropical tuna percentage (TTP) in longliner catches and the explanatory variable is the sea surface temperature (SST). This dataset consists of 77 observations of longliner catches in different points of the southern Indian Ocean in the year 2000 and one of these observations (observation 46) equals one, which goes to say that in that specific catch, only tuna was caught, which is a very unusual outcome compared to the rest of the data as the second highest value is 35%, for instance.

The methods used to fit a model for the analysis of this dataset were:

(1) Using the linear transformation proposed by Smithson and Verkuilen (2006) and then fitting a beta regression (BR) model via maximum likelihood and the robust approach (RobBR) proposed by Ribeiro and Ferrari (2020).

(2) Subtrating 0.01 from the boundary observation and then fitting a beta regression (BR) model via maximum likelihood and the robust approach (RobBR) proposed by Ribeiro and Ferrari (2020).

(3) Subtrating 0.001 from the boundary observation and then fitting a beta regression (BR) model via maximum likelihood and the robust approach (RobBR) proposed by Ribeiro and Ferrari (2020).

(4) Replacing observation 46 with the largest observation observed in the (0,1) interval in the dataset and and then fitting a beta regression (BR) model via maximum likelihood and the robust approach (RobBR) proposed by Ribeiro and Ferrari (2020).

(5) Removing observation 46 and and then fitting a beta regression (BR) model via maximum likelihood and the robust approach (RobBR) proposed by Ribeiro and Ferrari (2020).

(6) Fitting a one-inflated-beta regression model.

(7) Fitting a quasi-likelihood model proposed by Bonat *et al.* (2019) with an estimated p and with p fixed as 1.

The beta regression model used in methods (1) to (5) was fitted assuming that μ_i, which is the mean proportion of tuna caught in the ith fishing trip, and ϕ_i, the precision parameter for the ith observation, are

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 \times \text{SST}_i \tag{3.3}$$

and

$$\log(\phi_i) = \gamma, \tag{3.4}$$

respectively.

Method (6) was done fitting a one-inflated-beta regression model, with assumptions for μ_i and ϕ_i as in (3.3) and (3.4), with the distinction that μ_i is the conditional mean of y_i given that it is contained within the (0,1) interval. There is also the unknown mixture paramter α, which is assumed constant across all observations, hence $\alpha_i = \alpha \; \forall i = 1, \ldots, n$.

The quasi-likelihood model was fitted with the same assumption for μ_i as in the other models, but with the assumption that the variance of each observation is as stated in Section 2.2.2 where there is a power parameter p to be estimated.

After using methods (1) (applying linear transformation to the data), (2) (subtracting 0.01 from obs. 46) and (3) (subtracting 0.001 from obs. 46) and estimating the parameters via maximum likelihood note that observation 46 is considered to be influent on the model estimation and to be a leverage point (Figures A.1, A.2 and A.3), not to mention the models being considered a poor fit (Figures A.6a-c and Figures 3.3a-c). In each of these three situations, the BR model was fitted after using methods that, despite transforming the response variable so that it fits in the (0,1) interval, do not reduce the disparity between observation 46, which has response 1, and the rest of the observations. Hence, certain characteristics are similar in the diagnostics of these three models, which is shown in greater detail in Appendix A.

The models in methods (4) and (5) were fitted after using methods to transform the data that are more "aggressive", as when the boundary observation is replaced with the largest observation in (0,1) the method changes it to one that is within the range of the

rest of the data and in method (5) the observation is removed altogether. Because of this, the Cook's distance measure (see Appendix A) for observation 46 in both these models does not indicate that this observation is influential. Also, the worm plots in Figures 3.3d and 3.3e reinforce that these models are adequately fitted. The drawbacks of using such aggressive methods is that one is actually discarding information which may in fact be important when removing the observation and completely altering the response when replacing it, which is usually ill advised.

Table 3.1 shows the estimates for β_0, β_1 and ϕ in all fitted models. All estimates for β_1 are considered highly significant in all models and Figure 3.1 shows the fitted curves on a scatter plot with the observations from the data. Note that the curves respective to the three methods in which observation 46 was classed as influential appear above the rest of the curves, therefore in this particular scenario the boundary observation greatly affects how the regression model fits in the data. With regards to the models in methods (4) (replacing obs.46 by the highest observation in (0,1)) and (5) (removing obs. 46), note that they appear to be very close to the curve of the quasi-likelihood model for lower values of TTP, but the inclusion of the original observation 46 in the estimation of the model affects the gradient of the curve, making it elevate more than methods (4) and (5) as TTP increases. This effect is better observed in Figure 3.2, which presents a comparison between the fitted model 7 and a model using the same method without observation 46 in the dataset. Once more, the gradient of the curve of method (7) is greater than when fitting a model with a removed (or transformed) boundary observation.

Unlike other fitted models where observation 46 remained discrepant thus greatly affecting the estimates of the parameters, the robust beta regression puts less weight on the outlier thereby maintaining estimates close to the BR model fitted after removing the boundary observation while not excluding an observation which can provide information about the situation being studied. Note that when resorting to this type of estimation process for the regression model, subtracting 0.01 or 0.001 from the boundary observation hardly impacts the estimates of the parameters, hence any interpretation that can be made about the response and how the explanatory variable affects it remain the same regardless of the ϵ chosen to transform the boundary observation. Also note that using the robust estimation method does not impact estimates of models where the observations are not discrepant, indeed the estimates of the models fitted after replacing or removing the boundary observations are the same. This is due to the fact that the tuning constant q remains one in both these situations, whereas its value is 0.94 when subtracting 0.01 from observation 46 and 0.96 when subtracting 0.001 or when applying the linear transformation. The adequacy of these methods for this dataset is corroborated by the worm plots in Figures 3.3f, 3.4a and 3.4b, as well as the normal probability plots in Figures A.6f, A.6g and A.6h, where despite the residual value for obs. 46 being even a higher than when fitting a BR model, this is not an indicator of a poor fit as the robust estimation method is supposed to attribute a smaller weight to the discrepant observation, thereby making it so it does not have great impact on the estimates, which in turn results in an even higher residual value for obs. 46.

Even though an inflated-beta regression model was fitted and it provides a good fit, as per Figure 3.4e, it begs the question of whether it is worth adding an additional parameter α and changing the interpretation of μ_i in situations such as the one presented where there

is only one boundary observation and it does not hold any special meaning either. In this dataset, an observation equal to one means that in a particular day, all of the fish that were caught were tuna, which is a feasible occurrence, even if it is unlikely, given the percentage of tuna caught in the rest of the fishing trips recorded in the data. Also, catching only tuna does not have any special meaning, since it does not necessarily mean that there are not any other fish species that inhabit the area where the fishing took place. Therefore, I would argue that it is not worth fitting this type of model in scenarios similar to this one. Note that Figure 3.4e was created using the correct expression for the r_q residual, since there is an error with the residual calculation in the `gamlss` package. More details on the error and the solution to fix it are given in Section 3.2.

An advantage of using the quasi-likelihood method with $p = 1$ (compared to the BR models) in this scenario is that it is less affected by including observation 46 in the estimation process than the beta regression model (fitted by maximum likelihood) is affected by transforming the boundary observation to a close number such as when 0.01 or 0.001 is subtracted from it, therefore the method attributes less importance on the boundary observations when it comes to estimating the parameters. Nevertheless, the results are still slightly affected and the BR models estimated with the robust approach provide results closer to what they would be if observations 46 was excluded, not to mention the fact that the estimated standard errors of the estimates are significantly higher than the standard errors in the robust beta regression models. In the quasi likelihood model fitted estimating p, the power parameter p was estimated as 2.406 when observation 46 is present in the data, which would indicate that the distribution of this data differs to that of the beta distribution, where this parameter would be equal to one. However, when the boundary observation is excluded, the estimate of p is 1.071, which is close to one, resulting in a "variance function" very similar to that of the beta distribution. So even though the fitted curve is affected less than those of the methods which keep observation 46 as discrepant compared to when it is removed, the estimations are off the mark because the estimation method considers that the outlier changes the distribution of the response variable significantly, therefore the robust method is more conservative when dealing with an observation so different from the rest. Another disadvantage of this method in this scenario is that due to the estimate of p being so affected by the boundary observation, is that the dispersion is estimated as 1.88, which is larger than one, therefore it is not comparable to the precision parameter in the beta regression model as this would require the dispersion to be one at the very most, therefore analytically, it is impossible to put the dispersion estimate on the same scale as the γ shown in the rest of the models, hence why estimate of γ in the QL model is not shown in Table 3.1.

It is important to note that this dataset presents a very specific scenario where there is only one boundary observation and it differs greatly from the rest of the data. In order to prevent this outlier from greatly affecting the model estimates, consequently affecting the interpretation of the relationship between the response variable and the explanatory variable, the robust beta regression model is the best for this scenario, since it does not discard the information provided by the outlying observation, but bestows less importance on it compared to the usual beta regression model, thus mitigating its effect on the estimates. Since the boundary observation is so separated from the data, compressing all observations toward the centre of the unit interval with the linear transformation used in Smithson

and Verkuilen (2006) is not the most adequate solution, as even when using the robust estimation method proposed by Ribeiro and Ferrari (2020), the estimates are affected, even if not massively. Therefore, it is more advisable to simply subtract an $\varepsilon > 0$ from the boundary observation before proceeding with the model. As shown in Table 3.1 it does not matter whether the ε chosen is 0.01, 0.001 or perhaps another small value as the estimates remain approximately the same regardless of the choice.

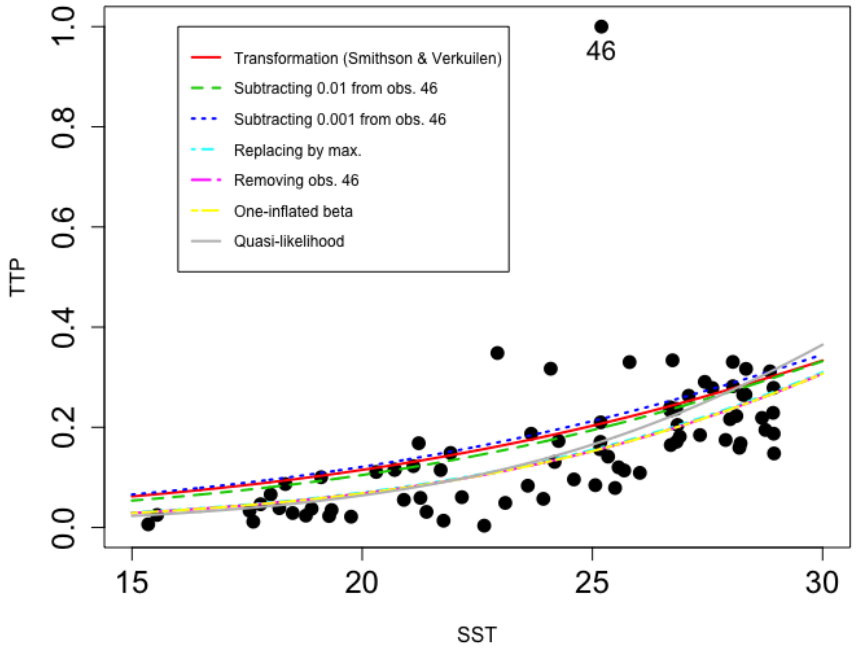


Figure 3.1: *Scatter graph of the data with the fitted curves for the models described in methods (1) to (7).*

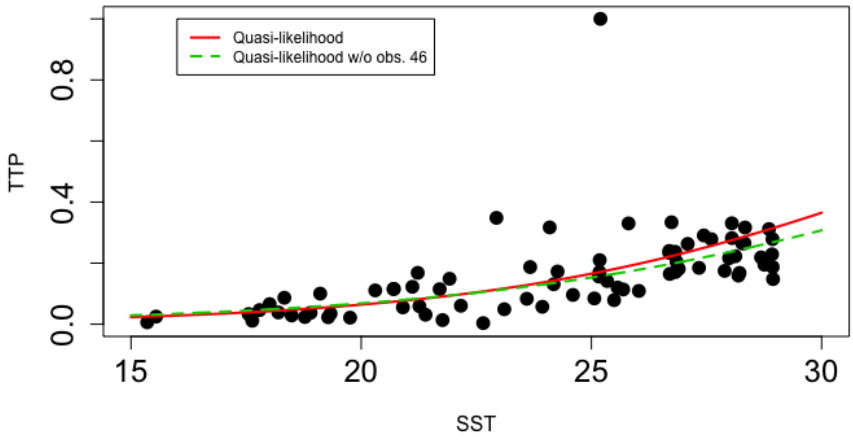
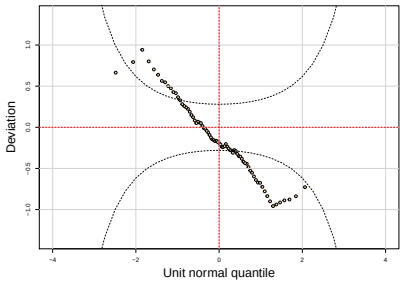


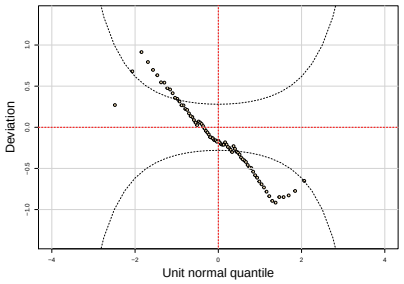
Figure 3.2: *Scatter graph of the data with the fitted curves for the quasi-likelihood model and a quasi-likelihood model fitted without observation 46 in the data.*

Table 3.1: *Point estimates for the mean, precision and additional parameters for the fitted models, along with their respective standard errors (between parentheses), in the tuna application.*

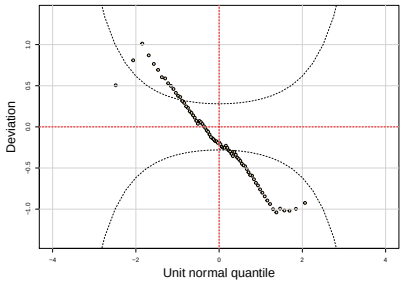
	Mean		Precision	Additional	
Model	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\gamma}$	$\widehat{\alpha}$	$\widehat{p}$
BR after using linear transformation	−4.737 (0.647)	0.135 (0.026)	2.030 (0.161)	-	-
BR after subtracting 0.01 from obs. 46	−5.039 (0.654)	0.145 (0.026)	2.052 (0.162)	-	-
BR after subtracting 0.001 from obs. 46	−4.666 (0.698)	0.134 (0.028)	1.723 (0.161)	-	-
BR after replacing obs. 46 with highest obs. in (0,1)	−6.162 (0.480)	0.179 (0.019)	3.239 (0.162)	-	-
BR without obs. 46 in the dataset	−6.223 (0.473)	0.180 (0.018)	3.300 (0.1635)	-	-
RobBR after using SV transformation	−5.884 (0.438)	0.169 (0.017)	3.413 (0.162)	-	-
RobBR after subtracting 0.01 from obs. 46	-6.206 (0.468)	0.180 (0.018)	3.330 (0.163)	-	-
RobBR after subtracting 0.001 from obs. 46	−6.207 (0.470)	0.180 (0.018)	3.313 (0.163)	-	-
RobBR after replacing obs. 46 with highest obs. in (0,1)	−6.162 (0.480)	0.179 (0.019)	3.239 (0.162)	-	-
RobBR without obs. 46 in the dataset	−6.223 (0.473)	0.180 (0.018)	3.300 (0.164)	-	-
One-inflated beta regression	−6.223 (0.470)	0.180 (0.018)	3.300 (0.163)	0.013 (0.013)	-
Quasi-likelihood regression	−6.938 (0.581)	0.213 (0.025)	*	-	2.406 (0.329)
Quasi-likelihood regression (p = 1)	−6.007 (0.808)	0.175 (0.031)	2.234 (0.722)	-	-



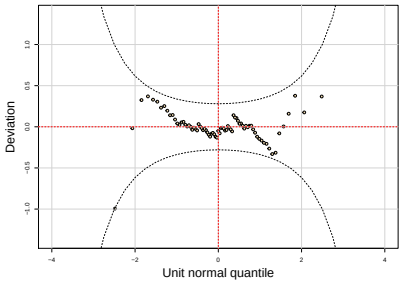
(a) *BR after linear transformation*



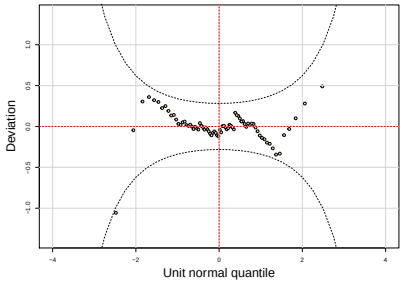
(b) *BR after subtracting 0.01 from obs. 46*



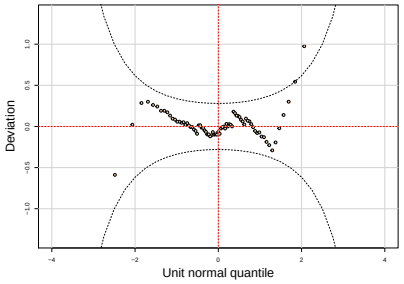
(c) *BR after subtracting 0.001 from obs. 46*



(d) *BR after replacing obs. 46 with largest obs. in (0,1)*

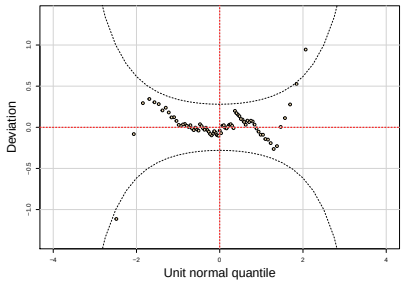


(e) *BR after removing obs. 46*

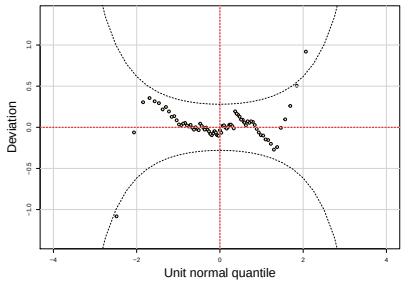


(f) *RobBR after linear transformation*

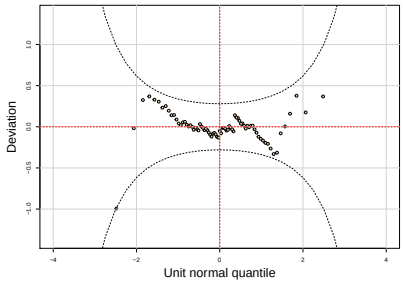
Figure 3.3: *Worm plots for the beta regression models and the inflated beta regression model in the tuna application.*



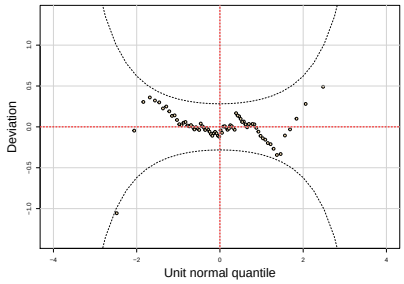
(a) *RobBR after subtracting 0.01 from obs. 46*



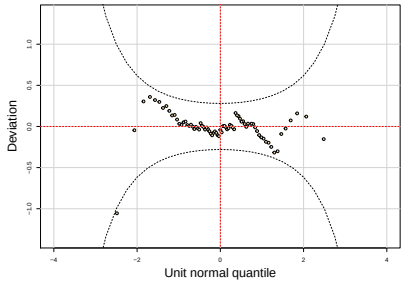
(b) *RobBR after subtracting 0.001 from obs. 46*



(c) *RobBR after replacing obs. 46 with largest obs. in (0,1)*



(d) *RobBR after removing obs. 46*



(e) *Inflated beta regression*

Figure 3.4: *Worm plots for the beta regression models and the inflated beta regression model in the tuna application.*

3.2 CVE application

Morrison *et al.* (2020) analyse the increase in vaccination exemptions in the United States. The authors state that the number of exemptions is increasing due to factors such as an increasing distrust in medical establishments, pervasive misinformation concerning vaccinations and declining health literacy regarding the possible severity of vaccine-preventable diseases. Declining levels of vaccine coverage are a consequence of this increase in vaccination exemptions, whether it be due to healthcare-related factors or vaccine hesitancy, and may result in a reemergence of measles and other vaccine-preventable diseases. Clusters of vaccination exemptions in a community can compromise herd immunity as, in general, 96% to 99% of people in that community must be vaccinated to ensure this type of immunity occurs.

In the United States, all fifty states require vaccination for school attendance, but exceptions are made, for medical reasons, in all of them, however fifteen states allow parents to opt out of the vaccination requirements via a conscientious vaccination exemption (CVE). CVEs on the rise in the US pose a public health threat and with the purpose of better understanding what can influence a higher number of CVEs in a community, Morrison *et al.* (2020) analysed the CVE occurrence in the state of Texas, which is the only state that does not require those who choose to refuse vaccination to be educated on the risks of doing so. The aim was to evaluate how sociodemographic and financial factors influence the amount of CVEs in a particular area. As the data is in the unit interval, the authors fitted a beta regression model after transforming the response variable using the linear transformation.

The data presented contains information on the percentage of children in a school system that have a CVE. Positive correlations have been found between CVE percentages and the percentage of a population in that area that is white and college educated, and in line with this correlation, CVE percentage tends to be higher in private schools. The analysis shown in this work will focus on a condensed version of the data presented by Morrison *et al.* (2020), where each observation represents a county in the state of Texas and the response variable exhibits the median CVE percentage when considering all school systems in each county. In total, there are 235 counties and in 12 of them, the median CVE percentage is equal to zero, thus there are 12 boundary observations in the dataset (5% of the observations). Also, the analysis is made considering data relative to the 2017-2018 time period.

The percentages of people with CVEs in the data are very close to zero as shown in Table 3.2. Therefore it is necessary to add a smaller ϵ to the boundary observations than in the previous applications, when using methods that require such a transformation. Note that there is a large difference between the third quartile and the maximum value of the response variable, which could indicate that there are discrepant observations (in the response variable) and these may or may not influence the estimates of the models. Indeed, the boxplot in Figure 3.5 shows that there are observations considered to be outliers, especially the highest value, however it remains to be seen whether these discrepant observations greatly impact the estimates of the models or not. Although it is possible, that when taking the covariates into account, these observations may not be discrepant

after all.

Table 3.2: *Descriptive statistics for the median CVE percentages in all counties.*

Minimum	1st quartile	Median	Mean	3rd quartile	Maximum	Standard Deviation
0.00000	0.00390	0.00795	0.00996	0.01313	0.09550	0.00990

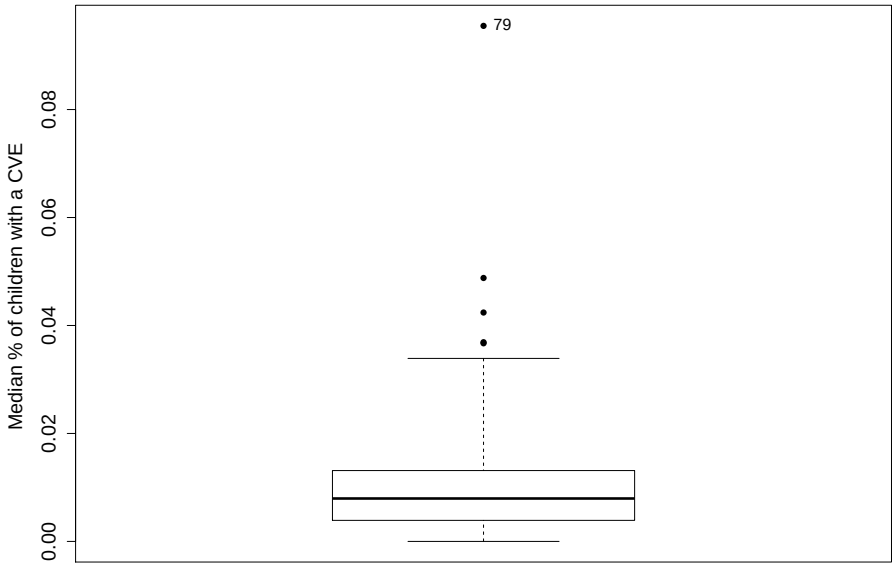


Figure 3.5: *Boxplot of the median CVE percentages in each county. Note that the boxplot is adjusted for the skewness of the data.*

After analysing the correlation between the response variable and the possible explanatory variables, then fitting some test models, the explanatory variables for the mean submodel were

- *% of the population in the county aged 5 and under who speak a language other than English at home* (ESL)
- *% of the population aged 25 and older in the county that have a bachelor's degree* (Bachelors)
- *% of people in that country whose race has been declared as white* (White)
- *Median household income in the county (dollars)* (Income).

The explanatory variables for the precision submodel were

- *% of children aged 5 and under in the county that are below the poverty line* (Poverty)
- *Dummy variable that has value 1 if the county is situated in a metropolitan area and 0 if it is not* (Metro)

The variable Poverty is not included in the mean submodel because it is masked by the variable Income, therefore when placing both in the model, Poverty is not classed as significant, however in the precision submodel Poverty is considered more significant than Income and that is the reason why they do not overlap in the submodels. The variable Metro, is not considered significant in the mean submodel, but is in the precision submodel, hence the fact that it is present in only one of the submodels.

The methods used to analyse this dataset were:

(1) Using the linear transformation proposed by Smithson and Verkuilen (2006) and then fitting a beta regression (BR) model via maximum likelihood

(2) Adding 0.0001 to the boundary observations and then fitting a beta regression (BR) model via maximum likelihood.

(3) Adding 0.00001 to the boundary observations and then fitting a beta regression (BR) model via maximum likelihood.

(4) Replacing boundary observations with the lowest observation observed in the (0,1) interval in the dataset and then fitting a beta regression (BR) model via maximum likelihood.

(5) Removing boundary observations and then fitting a beta regression (BR) model via maximum likelihood.

(6) Fitting a zero-inflated-beta regression model.

(7) Fitting a quasi-likelihood model proposed by Bonat *et al.* (2019) with an estimated p and with p fixed as 1.

The regression model used in methods (1) to (5) was fitted assuming that μ_i, which is the mean percentage of people that have a CVE certificate in the ith county, and ϕ_i the associated precision parameter are

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 \times \text{ESL}_i + \beta_2 \times \text{Bachelors}_i + \beta_3 \times \text{White}_i + \beta_4 \times \text{Income}_i \tag{3.5}$$

and

$$\log(\phi_i) = \gamma_0 + \gamma_1 \times \text{Poverty}_i + \gamma_2 \times \text{Metro}_i, \tag{3.6}$$

respectively.

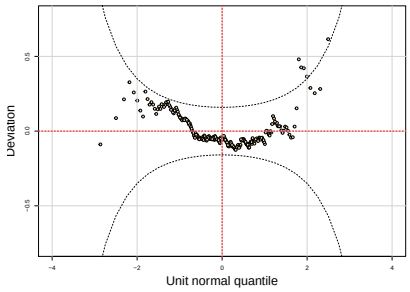
The zero-inflated beta regression assumes that the mean and precision are in the same form as shown in (3.5) and (3.6), however μ_i in that case, represents the mean percentage of people that have a CVE certificate in the ith county given that the observation is in (0,1). Also, the mixture parameter α represents the probability that the median percentage of children with a CVE in the ith county is zero. One of the test models fitted included an explanatory variable for α, however no noticeable improvement on the model's quality of fit was noticed and in this scenario, there arguably is not much of a difference between an observation equal to zero and an observation very close to zero as it does not indicate that there is anything peculiar happening in the county if a number close to zero is observed.

In fact, as explained by Morrison *et al.* (2020), if only an absolutely tiny percentage of people have a CVE in a community, there is not a public health threat associated with this value, therefore any interpretation of such a value would be very similar to an observation that is exactly zero. The quasi-likelihood model assumes that the mean is given by (3.6), but since the precision remains constant in that model, only γ_0 is estimated.

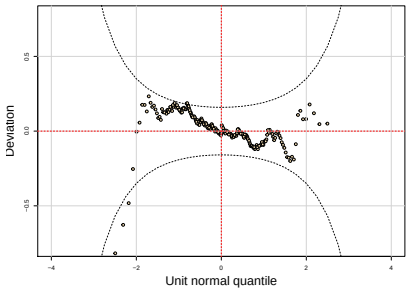
Figure 3.6 presents the worm plots for the fitted models. Note that Figure 3.6e shows that the model is well fitted when excluding the boundary observations from the dataset. Figure 3.6d also presents a decent fit when replacing the boundary observations with the smallest in the (0,1) interval, also note that this worm plot is similar to the worm plot for the model fitted after adding 0.0001 to the observations equal to zero. When using the linear transformation, the worm plot seems to indicate a worse fit than the aforementioned models as there are more observations near the delimited boundary, but the transformation still enables a better fit than adding 0.00001 to the boundaries (Figure 3.6d), as this value is very low, the model is being greatly affected, which is translated in the worm plot not indicating a good fit.

In Appendix B, there are more graphs that assist in the diagnostics for the fitted models. Figures B.1, B.2, B.3, B.4 and B.5 seem to indicate that the models are fairly well fitted. In the normal probability plots with simulated envelopes in Figure B.6, notice that the residuals are, mostly, inside the envelopes, however it is possible to see that the models in which the worm plots indicated were most well fitted (BR after excluding boundary observations, BR after replacing boundary observation with the smallest obs. in (0,1) and BR after adding 0.0001 to obs. equal to zero) have easily identifiable points that lie outside of the envelopes; they are few and close to the area inside the envelopes. However, in the rest of the models, these same points are not shown, which is due to them being farther from the boundaries to the point where they do not appear on the graph, indicating a worse fit. In each of the Figures B.1, B.2, B.3, B.4 and B.5, graph (b) shows Cook's distance measure for each of the observations in each of the models, note that observation 79 has a particularly high value. This observation has the highest response (shown in Table 3.2) and the graphs would indicate that it is a very influent observation and should be analysed further, however fitting the exact same models in the absence of this observation produces very similar estimates to the models already fitted in the presence of county 79. Therefore, despite its Cook's distance value, this observation is not affecting the estimates. The fact that it does not affect the estimates is also the reason why the robust estimation method for the beta regression model selects $q = 1$, thus being equivalent to estimating the beta regression model via maximum likelihood in this scenario. In other words, the RobBR model evaluates that it is not worth attributing a different weight to this discrepant observation.

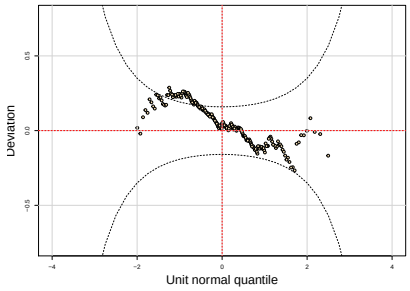
Table 3.3 presents the estimates for the parameters of the fitted models. Note that, as was the case in the worm plots, BR after adding 0.0001 to the boundary observations and BR after replacing observations equal to zero with the smallest obs. in (0,1) provide similar fits. This could be due to 0.0001 being close to the smallest obs in (0,1) (0.00025), however it could be argued that given the magnitude of the data, these values are not that close given 0.00025 is more than double 0.0001. Either way, the model that adds 0.0001 is not too affected by this difference, which is a positive takeaway. The same cannot be said about adding 0.00001 to the boundary observations, as the estimate of β_1 is particularly



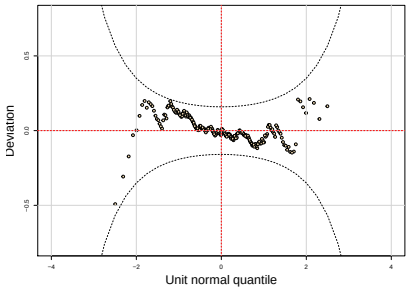
(a) *BR after linear transformation*



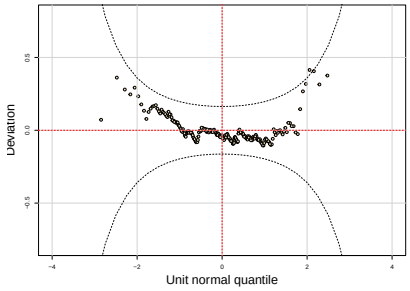
(b) *BR after adding 0.0001 to boundary obs.*



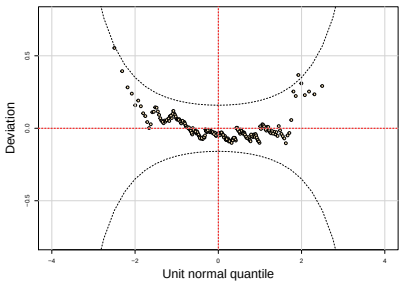
(c) *BR after adding 0.00001 to boundary observations*



(d) *BR after replacing boundary obs with smallest obs. in (0,1)*



(e) *BR after removing boundary observations*



(f) *Inflated beta regression*

Figure 3.6: *Worm plots from beta regression models and the inflated beta regression model fitted for the CVE data.*

affected, along with a significant reduction in the precision parameters. It may not seem to be a significant change in the estimates without taking context into account, but it is important to recall that the values of the response variable are very low, therefore a change in the estimates greatly impacts the residuals and the overall quality of the fit, which is corroborated by Figure 3.6d.

The BR model fitted after using the linear transformation produces estimates that are very discrepant from the other fitted models. Note that there is an approximate 0.2 difference from the intercept estimate when compared to the rest of the fits. These discrepant parameter estimates paired with the fact that the diagnostics previously discussed do not indicate a good fit for this model, suggest that this transformation is distorting the data to the point where the estimates become significantly different to the rest of the fits. This could be due to the magnitude of the observations, since they are very small and the

transformation essentially calculated a weighted average between the observation and 0.5, thereby making it a fairly aggressive transformation in this particular scenario. Figure 3.7 shows a comparison between the model fitted after using the linear transformation and other models which the diagnostics showed to be well fitted. Note how using the linear transformation resulted in fitted values that are systematically higher than those of the other models (apart from one observation at the end). As the other models are well fitted, it is worrying that the linear transformation affected the response value to the point were the fitted values alter so much. If the fitted values fluctuated between being higher and lower to the fitted values for the other models, it would be more acceptable, however the evidence suggests that the estimates for the BR model after using the linear transformation are quite distorted. These findings are expanded on in the second simulation scenario of Chapter 4, which is based on this application and the linear transformation also negatively affects the estimation of the BR model.

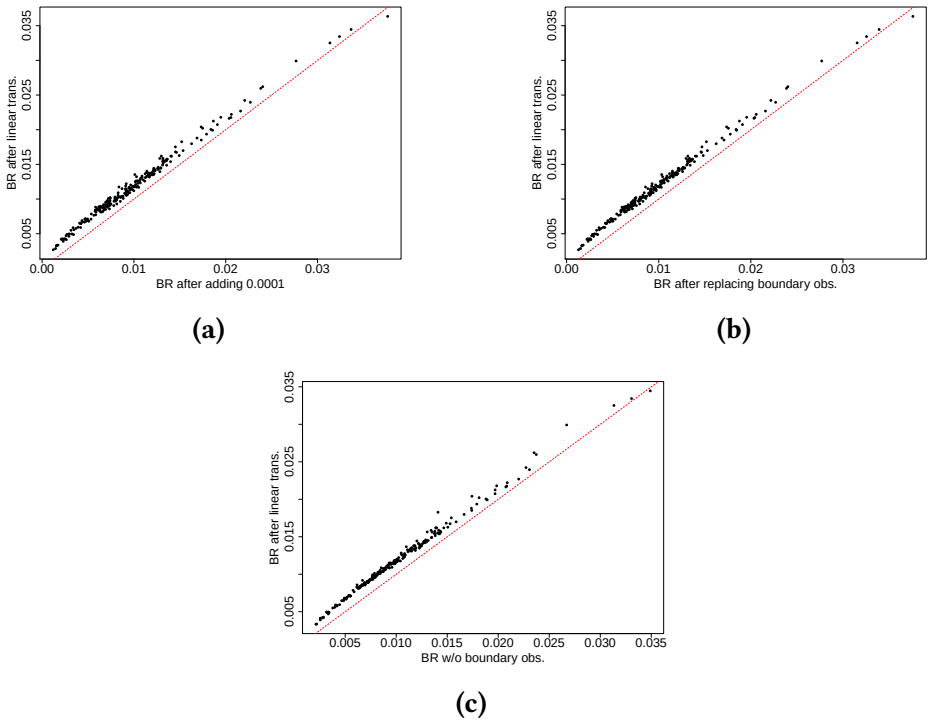


Figure 3.7: *Scatter plots comparing fitted values from the BR model after using the linear transformation with (a) the BR model fitted after adding 0.0001 to the boundary obs.; (b) the BR model fitted after replacing the boundary obs. with smallest in (0,1) and (c) the BR model fitted after excluding the boundary observations. The diagonal line in each figure is a line with intercept equal to zero and slope equal to one where all points would lie if the fitted values were equal in both models.*

The quasi-likelihood model while estimating p produces estimates that differ from the other models, especially when it comes to the estimate of β_1. Not considering the BR model fitted after using the linear transformation, the QL model's estimate of β_1 is less than half the estimate of β_1 in the rest of the models; given that the ESL variable is considered highly significant in the model, this very discrepant estimate raises questions about the estimation accuracy of the quasi-likelihood model in this application. Also, the estimation of γ_0 results in a very high standard error estimate, larger even than the point estimate of

γ_0, this could be due to the fact that this model is limited with regards to modelling the precision, as it needs to be constant. Note that Figure B.7 does not indicate that this model provides a bad fit for this data, but there is a surplus of residuals below zero in the left side of the plot when compared with residuals greater than zero, which could possibly raise questions as to how well fitted this model really is. The model estimates p as 0.767 with a standard error of 0.403, therefore if one were to create a 95% confidence interval for p, 1 would be included in the interval and that would mean the "variance function" in this model would be equivalent to that of the beta distribution. The QL model with p fixed as one also yields poor results when compared to well fitted models, the estimate for β_1 is less than half of the estimate in the well fitted models and the other estimates are far off as well.

The worm plot in Figure B.8, calculated with the default residuals from the `gamlss` package, would seem to suggest that the zero-inflated beta regression model provides a poor fit for this dataset. However, notice that the estimates shown in Table 3.3 of the parameters of the mean submodel of the zero-inflated BR model are equal to those of the BR model after excluding the boundary observations, therefore the model is doing exactly what it is supposed to do, which is estimating μ_i not considering the boundary observations. Given this dilemma, it is important to try and understand why the worm plot is displaying such a poor fit for the model and if the fit is in fact bad. Firstly, analysing the shape of the deviation points in the worm plot, it is clear that these points take a plunge in the highest values of the normal quantile, which suggests that in these observations, the residual value is getting lower and lower from what they should be if the randomized quantile residual (r_q) was distributed according to a normal distribution. Essentially, the worm plot is suggesting that these residuals are not following a normal distribution, when it should if the model was correct. To verify that this worm plot does not only suggest a poor fit in this particular sample, a simulation study was conducted, where nine different samples were randomly generated, in which each one had the same size as the orginal sample found in the dataset in analysis and for each sample $y'_1, \dots, y'_{239}$ it was assumed that

$$y'_i \sim \text{BEZI}(\widehat{\alpha}, \hat{\mu}_i, \widehat{\phi}_i),$$

where $\widehat{\alpha}$, $\hat{\mu}_i$ and $\widehat{\phi}_i$ refer to the fitted values for the ith observation of the original sample using the parameter estimates for the zero-inflated BR model displayed in Table 3.3. The explanatory variable values for each observation i were the same as in the respective ith observation in the CVE dataset.

After generating each sample randomly, a zero-inflated BR model was fitted. Figure B.9 shows the worm plots of the nine fitted models for their respective generated samples. Despite every model accurately estimating the pre-defined model parameters, note how the worm plots all show a similar pattern to the one in Figure B.8, where the observations that are in the highest quantiles of the normal distribution plunge downwards, indicating a bad fit. Therefore, it is possible to conclude that there is an issue with the residuals in this scenario that consequently affects the usefulness of the worm plot for this specific model in this particular case. In a similar simulation study where the sample size is quadrupled resulting in 956 observations, and with the explanatory variables being cloned three times

as to result in the same number of observations as the response variable, then proceeding in a similar way to the last study, note how the worm plots for the fitted models, shown in Figure B.10, again display a similar pattern to those previously discussed. Also, Figure B.11 which is a usual normal probability plot with the residuals of one of the models fitted in a generated sample of size 956, suggest that despite a large sample, the residuals do not seem to be converging to a normal distribution in the tails, especially in the upper tail. This confirms that the worm plot does not accurately depict whether the model is well fitted or not in this case, because the residuals themselves are not normally distributed. When exploring the reason for the residuals not to be normally distributed, it was found that the r_q residual for the inflated beta regression model is not correctly programmed on the `gamlss` package on `R`. The problem lies in the fact that when calculating the residual, rather than using the expression in (3.1) for the observations in (0,1), which uses the cumulative distribution function of the beta distribution, the distribution function of the zero-inflated beta distribution is being used instead, hence calculating $(1-\alpha)\text{BI}_0(y;\widehat{\alpha},\hat{\mu},\widehat{\phi})$, which would equate to

$$\widehat{\alpha}(1-\widehat{\alpha}) + (1-\widehat{\alpha})^2 F(y_i;\hat{\mu},\widehat{\phi}),$$

where $F(\cdot)$ is the distribution function of the beta distribution, thus resulting in a different expression compared to (3.1). The worm plot in Figure 3.6f was created using the correct expression for the r_q residual in the case of the inflated beta regression model and it confirms that the model is indeed well fitted. Also, with the nine simulated samples which previously showed incorrect worm plots, the r_q residual with the correct expression was calculated and Figure B.12 shows the worm plots created with the correct residual values, showing a good fit for the models. As was the case with the nine generated samples with quadruple sample size, shown in Figure B.13. The implementation of the correct expression for the randomized quantile residual can be found in https://github.com/danielanobrega/BR_boundaries.

Looking at the estimates of the models' parameters in Table 3.3, note that the initial hypothesis from Morrison *et al.* (2020) that there was a positive correlation between CVE percentage and being white and college educated is confirmed by the fitting of the models; the estimates for β_2 and β_3 are considered to be significant and are positive, thus as the percentage of people with these characteristics in the sample increase, the CVE percentage, on average, also increases. The estimate of β_1 is the only one that implies a reduction on the mean CVE percentage for the county. This is interesting as ESL tends to refer to immigrants, therefore they either do not share the same views on vaccination as the native population (especially those who are considered white and have a bachelor's degree) or because the immigrants are more likely to follow rules imposed by the government in order to avoid any sanctions which will affect their lives much more than families that have an American citizenship. Analysing the estimate of γ_2, one can infer that amongst metropolitan counties, the precision is around 277 times more precise if the value of Poverty is fixed, so there tends to be more condensed values for CVE percentages in counties situated in a metropolitan area.

To analyse this dataset, the BR model after adding 0.0001 to the boundaries and the zero-inflated beta regression model seem to be the best alternatives. Although the BR model after replacing the boundary observations with the smallest observed value in (0,1) and the BR model fitted without the observations equal to zero provide decent fits, the former makes it so observations to zero are the same as observations equal to 0.00025

which distorts the data. Whereas, adding 0.0001 to the boundary observations results in a similar fit and this method still differentiates the response values equal to zero from all the others, while still maintaining a decent fit, unlike the method where 0.00001 is added, which seems to cross a threshold where the model is highly affected by such low values. With regards to excluding the boundary observations, these constitute 5% of the data, thus a significant proportion of the dataset will be discarded, which is not ideal. However, the inflated beta regression model maintains the same estimates for the mean submodel in (0,1) and does not discard observations that may bring important information to the analysis, thus being a good alternative for this scenario. Even though it can be argued that this specific data does not have a particular and special reason to separate the boundary observations, it also does not have a reason that impedes this model's use in a way that it would not make sense to separate the data. Therefore, I would argue that it is an option and one can decide between separating the analysis and choosing to fit the inflated beta regression model or not separating and adding 0.0001 to the boundary observations.

Table 3.3: *Point estimates for the rmean, precision and additional parameters for the fitted models, along with their respective standard errors (between parentheses) and p-values [between brackets].*

	Mean					Precision			Additional	
Model	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\gamma}_0$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\alpha}$	$\widehat{p}$
BR after using linear transformation	−4.527 (0.032) [<0.001]	−0.283 (0.033) [<0.001]	0.183 (0.039) [<0.001]	0.111 (0.032) [<0.001]	0.132 (0.039) [<0.001]	5.637 (0.101) [<0.001]	0.639 (0.095) [<0.001]	0.841 (0.229) [<0.001]	-	-
BR after adding 0.0001 to boundary obs.	−4.771 (0.048) [<0.001]	−0.418 (0.048) [<0.001]	0.209 (0.050) [<0.001]	0.113 (0.042) [0.007]	0.149 (0.051) [0.003]	5.115 (0.104) [<0.001]	0.487 (0.093) [<0.001]	1.019 (0.225) [<0.001]	-	-
BR after adding 0.00001 to boundary obs.	−4.773 (0.052) [<0.001]	−0.447 (0.051) [<0.001]	0.208 (0.052) [<0.001]	0.096 (0.044) [0.028]	0.142 (0.054) [0.009]	4.918 (0.106) [<0.001]	0.433 (0.092) [<0.001]	1.112 (0.224) [<0.001]	-	-
BR after replacing boundary obs. with lowest obs. in (0,1)	−4.769 (0.046) [<0.001]	−0.403 (0.046) [<0.001]	0.208 (0.049) [<0.001]	0.121 (0.041) [0.003]	0.153 (0.050) [0.002]	5.206 (0.104) [<0.001]	0.517 (0.093) [<0.001]	0.973 (0.225) [<0.001]	-	-
BR without boundary obs. in the dataset	−4.727 (0.044) [<0.001]	−0.356 (0.048) [<0.001]	0.203 (0.046) [<0.001]	0.155 (0.040) [<0.001]	0.148 (0.047) [0.002]	5.410 (0.107) [<0.001]	0.418 (0.101) [<0.001]	0.763 (0.231) [<0.001]	-	-
Zero-inflated beta regression	−4.726 (0.045) [<0.001]	−0.356 (0.049) [<0.001]	0.203 (0.048) [<0.001]	0.155 (0.040) [<0.001]	0.148 (0.050) [0.003]	5.409 (0.109) [<0.001]	0.418 (0.107) [<0.001]	0.763 (0.251) [0.003]	0.050 (0.014) [<0.001]	-
Quasi-likelihood regression	−4.695 (0.066) [<0.001]	−0.174 (0.076) [0.022]	0.158 (0.069) [0.021]	0.198 (0.067) [0.003]	0.180 (0.070) [0.010]	5.889 (22.896) [0.509]	-	-	-	0.767 (0.403)
Quasi-likelihood regression (p = 1)	-4.703 (0.066)	−0.194 (0.076)	0.151 (0.072)	0.199 (0.067)	0.196 (0.073)	4.770 (0.501)	-	-	-	-

Chapter 4

Simulation

To evaluate the performance of the methods discussed in Chapter 2, Monte Carlo simulations were conducted considering different scenarios. For each of the scenarios, N = 10000 replicates were done.

To evaluate the performance of the methods in estimating the parameters, the estimated bias and the root mean squared error (RMSE) measures were used. Let θ represent a parameter, $\widehat{\theta}$ the estimator for this parameter and $\widehat{\theta}_j$ is the estimate of θ in the jth replicate of the simulation. The estimated bias and estimated RMSE are

$$\widehat{\text{Bias}}(\widehat{\theta}) = \frac{1}{N}\sum_{j=1}^{N}\widehat{\theta}_j - \theta$$

and

$$\widehat{\text{RMSE}}(\widehat{\theta}) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\widehat{\theta}_j - \theta\right)^2}.$$

4.1 First scenario

The first scenario simulated is one where there is a lower detection limit to the response variable y. In such situations, any observation below a certain value is registered as zero despite it not being a true zero. The sample sizes considered in each replicate were n = 50, 100. One explanatory variable was considered for this example, which was generated with random draws from a standard uniform distribution, i.e.

$$x_i \sim U(0,1),\ \ i = 1,\ldots,n,$$

where $U(\cdot)$ denotes the uniform distribution. The covariate x was set for the sample size of 50 and then cloned to obtain the values corresponding to the sample size of 100, therefore $x_i = x_{i+50}, i = 1,\ldots,50$, in that sample. The covariate x was generated once, before the simulation, and maintained the same across all replicates of the simulation. The logit and logarithmic functions were used as link functions for the mean and precision, respectively.

Hence, μ_i and ϕ_i are

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_i \tag{4.1}$$

and

$$\log(\phi_i) = \gamma, \tag{4.2}$$

In every replicate of the simulation, two samples of the response variable were randomly generated taking random draws from a beta distribution, so that

$$y_i \sim \mathcal{B}(\mu_i, \phi_i), \;\; i = 1, \ldots, n,$$

where μ_i and ϕ_i are obtained from applying the inverse logit function and the exponential function to (4.1) and (4.2), respectively. The parameter values were defined as $\beta_0 = -3.7$, $\beta_1 = 1.6$ and $\phi \equiv \exp(\gamma) = \exp(4.094345) = 60$. With these values, the linear correlation coefficient between y and x across the 10000 replicates varies from 0.389 to 0.798 with 75% of the values being higher than 0.604 in the sample size of 100 and in the sample size of 50, the correlation varies from 0.195 to 0.879 with 75% of the values being higher than 0.595. These values for the parameters resulted in values for μ that varied from 0.024 to 0.108 with the median value being 0.043.

After generating the samples, in order to simulate the detection limit scenario, the observations that were lower then 0.5% were registered as zero. There will be samples where no zeros will be present in the sample, so the methods used had to be adapted. If there was at least one zero in the sample, the fitted models were

(1) Using the linear transformation proposed by Smithson and Verkuilen (2006) and then fitting a beta regression (BR) model via maximum likelihood.

(2) Adding 0.005 to the boundary observations and then fitting a beta regression (BR) model via maximum likelihood.

(3) Adding 0.0005 to the boundary observations and then fitting a beta regression (BR) model via maximum likelihood.

(4) Replacing boundary observations with the lowest observation observed in the (0,1) interval in the dataset and then fitting a beta regression (BR) model via maximum likelihood.

(5) Removing boundary observations and then fitting a beta regression (BR) model via maximum likelihood.

(6) Fitting a zero-inflated-beta regression (BR) model via maximum likelihood.

(7) Fitting a quasi-likelihood model proposed by Bonat *et al.* (2019) while estimating p and with p fixed as 1.

If there were no zeros in the sample, methods (1) to (6) were changed to a beta regression model without using any transformations to map the data onto (0,1). Therefore, in the replicates in which there were no zeros in the sample, methods (1) to (6) were equivalent to each other, thus producing the same estimates. Figure 4.1 presents how many zeros were in each sample across the 10000 replicates of the simulation. Note that in the sample

size 100, less than a thousand replicates had samples where there were no zeros and few replicated had samples with more than 5 zeros. In the sample size of 50, the number of samples with no zeros increases due to the smaller sample size reducing the amount of observations lower than 0.5% in the sample and very few replicates had samples with more than 3 zeros.

In this scenario, the robust estimation in the beta regression model yields the same results as the maximum likelihood estimation seeing as there are not discrepant observations that greatly influence the estimates. Therefore, this robust method will not be analysed as any findings would be equivalent to the usual estimation in BR models.

The quasi-likelihood model where p is not previously defined presented problems in some of the replicates. Depending on the generated sample, the model's Cholesky factorization fails, thus resulting in a failure in the model's estimation. There does not seem to be a particular pattern among the samples that resulted in failure. Out of the 10000 replicates, the QL model failed in 3188 for the sample with 50 observations and in 2568 replicates for the sample size of 100. Therefore, the estimates from this model are not shown in the table.

Table 4.1 presents the results of the estimated bias and estimated RMSE for the models fitted in the simulation. The BR model fitted after using the linear transformation results in poor estimates for the parameters when compared to the other methods used prior to fitting a beta regression model, note that the bias for the estimates is significantly higher than in the other BR models. This is in accordance with the CVE application where the observation values were small and the linear transformation affected the fitted model. This reinforces the point stated in the analysis of the CVE application that the use of this transformation must not be automatic, as it often is. It is necessary to analyse the circumstances of the data in order to decide what the best method to map the data onto (0,1) is.

When fitting the QL model with p fixed as one, the results are very good. The advantage of resorting to this model is that there is no need to transform the data before using it, therefore it avoids a certain amount of subjectivity when deciding which method to transform data with boundary observations in order to fit a beta regression model.

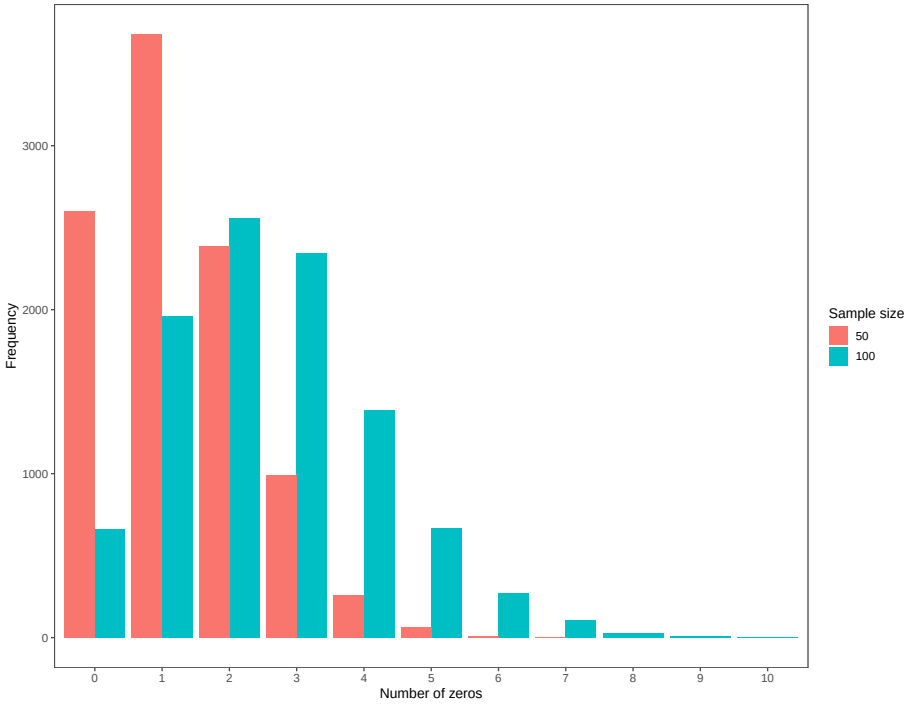


Figure 4.1: *Bar plot of the number of zeros in each sample for every replicate of the simulation.*

Table 4.1: *Estimated bias and root mean squared error (RMSE) for the estimates of the parameters when the sample size is 50 and 100 in all replicates of the first simulation scenario.*

		$n = 50$		$n = 100$	
Model		Bias	RMSE	Bias	RMSE
BR after using linear transformation	β_0	0.238	0.291	0.155	0.187
	β_1	−0.210	0.312	−0.137	0.215
	γ	0.274	0.351	0.159	0.216
BR after adding 0.005 to boundary obs.	β_0	0.006	0.163	0.012	0.114
	β_1	−0.018	0.256	−0.024	0.180
	γ	0.102	0.230	0.069	0.160
BR after adding 0.0005 to boundary obs.	β_0	−0.050	0.189	−0.046	0.136
	β_1	0.074	0.290	0.071	0.210
	γ	−0.035	0.236	−0.073	0.178
BR after replacing boundary obs. with lowest obs. in (0,1)	β_0	0.016	0.164	0.018	0.115
	β_1	−0.034	0.258	−0.033	0.182
	γ	0.122	0.242	0.081	0.166
BR without boundary obs. in the dataset	β_0	0.068	0.175	0.073	0.134
	β_1	−0.103	0.275	−0.109	0.210
	γ	0.146	0.256	0.111	0.185
Zero-inflated beta regression	β_0	0.068	0.175	0.073	0.134
	β_1	−0.103	0.275	−0.109	0.210
	γ	0.146	0.256	0.111	0.185
	α	0.026	0.034	0.026	0.030
Quasi-likelihood regression ($p = 1$)	β_0	−0.015	0.181	−0.011	0.127
	β_1	0.013	0.283	0.011	0.200
	γ	0.067	0.246	0.039	0.171

4.2 Second scenario

The second simulation scenario was based on the CVE application in Section 3.2. To generate the data, the covariates from the CVE dataset were used, thus, unlike the previous scenario, the explanatory variables were not randomly generated.

The logit and logarithmic functions were used as link functions for the mean and precision, respectively. Hence, μ_i and ϕ_i are as shown in (3.5) and (3.6), respectively.

In every replicate of the simulation, a response variable was randomly generated taking random draws from a beta distribution, so that

$$y_i \sim \mathcal{B}(\mu_i, \phi_i), \;\; i = 1, \ldots, n,$$

where μ_i and ϕ_i are obtained from applying the inverse logit function and the exponential function to (3.5) and (3.6), respectively. The parameter values were fixed as the parameter estimates from the BR model fitted after adding 0.0001 to the boundary observation presented in Table 3.3 as this model was concluded to be well fitted and its use was recommended for the analysis of that data. Therefore the fixed parameter values were $\beta_0 = -4.771$, $\beta_1 = -0.418$, $\beta_2 = 0.209$, $\beta_3 = 0.113$, $\beta_4 = 0.149$, $\gamma_0 = 5.115$, $\gamma_1 = 0.487$ and $\gamma_2 = 1.019$. With these parameter values, μ varies from 0.0012 to 0.0375 with the median value being 0.0092.

After generating the sample, as was the case in the first scenario, a detection limit value was fixed. In this case, however, any observation of the response variable whose value was below 0.0001 was changed to zero. In this situation, there were considerably less replicates in which the generated sample had no zeros than in the first scenario. If there was at least one zero in the sample, the fitted models were

(1) Using the linear transformation proposed by Smithson and Verkuilen (2006) and then fitting a beta regression (BR) model via maximum likelihood.

(2) Adding 0.0001 to the boundary observations and then fitting a beta regression (BR) model via maximum likelihood.

(3) Adding 0.00001 to the boundary observations and then fitting a beta regression (BR) model via maximum likelihood.

(4) Replacing boundary observations with the lowest observation observed in the (0,1) interval in the dataset and then fitting a beta regression (BR) model via maximum likelihood.

(5) Removing boundary observations and then fitting a beta regression (BR) model via maximum likelihood.

(6) Fitting a zero-inflated-beta regression model.

(7) Fitting a quasi-likelihood model proposed by Bonat *et al.* (2019).

If there were no zeros in the sample, methods (1) to (6) were changed to a beta regression model without using any transformations to map the data onto (0,1). Figure 4.2 shows the

majority of the generated samples had 3-6 zeros in the sample, which would represent 1.26% to 2.5% of the data, thus being a very low percentage of the data. There were no zeros in the sample in only 46 replicates, therefore in most of the 10000 replicates, methods (1) to (6) yielded different results.

This scenario is also one in which the robust estimation method for beta regression model yields the same results as the maximum likelihood estimation therefore the results do not focus on this robust method of estimation. Note that even in the CVE application, which this simulation is based on, the robust beta regression model did not consider any observations to be greatly influent on the estimates.

As stated in Section 3.2, this particular QL model is only implemented for a fixed precision, therefore the model will only estimate γ_0. As was the case in the previous scenarios, there were also replicates in which the QL model did not work, however in this case, this issue did not occur as often as in the previous simulation; the QL model did not work in 502 replicates in this scenario. Even still, the results for the QL model while estimating p are not present in this section.

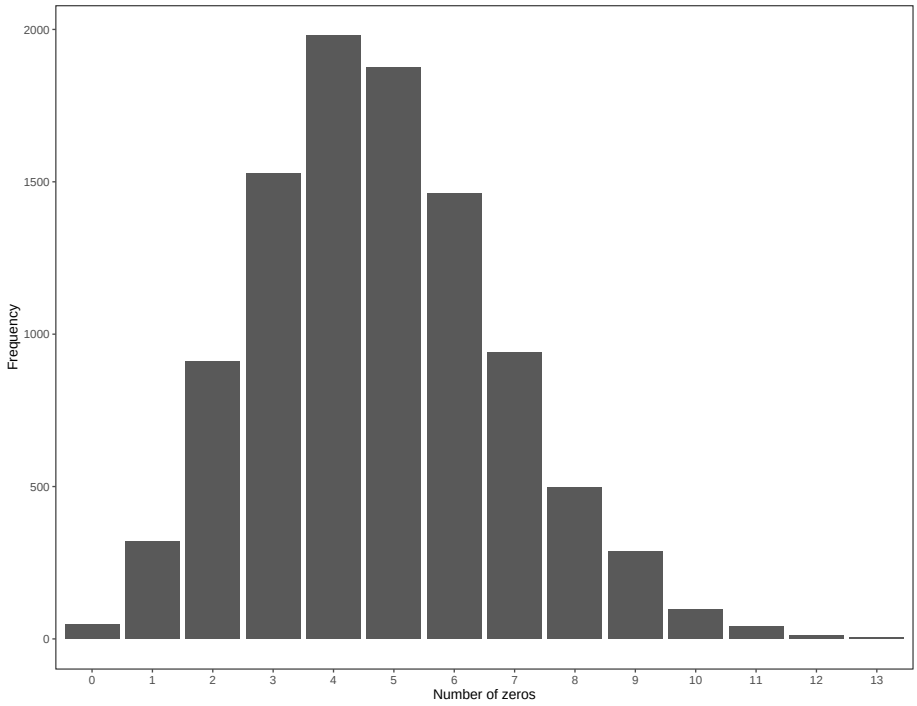


Figure 4.2: *Bar plot of the number of zeros in each generated sample for all 10000 replicates in the second simulation scenario.*

Table 4.2 presents the results of the simulation. In the CVE application's findings, the poor performance of the BR model fitted after using the linear transformation was commented on and combined with the first simulation scenario, there seems to be an indication that with very small values of the response variable, the transformation seem to greatly affect the estimates of the subsequently fitted BR model. The results for this method in this simulation scenario show a very high bias in the estimates of β_0, β_1 and γ_0 especially. These biases are particularly alarming when one considers that due to the low values of the observations, even a small deviation from the correct parameter values result

in a massive change in the estimated value for the response variable for each observation in the sample. Therefore, this emphasises that the method of mapping the data onto (0,1) does not seem to work well when analysing the CVE data, which was the impression when interpreting the results in Section 3.2.

With regards to the quasi-likelihood model, note that it has satisfactory results when it comes to the estimated bias, but the variability of the estimates is higher than the other BR models (aside from the BR model after using the linear transformation). However, this is a situation where it would be necessary to have covariates in the precision submodel, thus the use of the QL model is not completely adequate, since the model is limited to a fixed precision.

The inflated beta regression model seems to have had a worse performance than the methods in which an ε was added to the boundary observations that preceded the fitting of a BR model, but this may be due to the fact that the generated data follows a beta distribution and not a zero-inflated beta distribution, furthermore the model is attempting to estimate the parameters with fewer observations than the BR models, since the estimation is done putting aside the boundary observations. The same logic applies to the BR model fitted after removing the boundary observations. The two best models in this scenario, which were the two in which an ε was added to the boundary observations, performed similarly.

Table 4.2: *Estimated bias and root mean squared error (RMSE) for the estimates of the parameters in the second simulation scenario.*

		Mean					Precision			Additional
Model		β_0	β_1	β_2	β_3	β_4	γ_0	γ_1	γ_2	α
BR after using linear transformation	Bias	0.2556	0.1469	−0.0417	−0.0153	−0.0240	0.4587	0.0405	−0.1423	
	RMSE	0.2586	0.1503	0.0577	0.0357	0.0475	0.4716	0.1069	0.2725	-
BR after adding 0.0001 to boundary obs.	Bias	0.0008	0.0158	−0.0028	0.0009	0.0017	0.0594	−0.0055	0.0082	
	RMSE	0.0479	0.0488	0.0496	0.0412	0.0512	0.1198	0.0945	0.2365	-
BR after adding 0.00001 to boundary obs.	Bias	-0.0055	−0.0030	0.0018	0.0002	−0.0012	0.0140	−0.0006	0.0600	
	RMSE	0.0489	0.0493	0.0505	0.0425	0.0522	0.1080	0.0997	0.2484	-
BR after replacing boundary obs. with lowest obs. in (0,1)	Bias	0.0017	0.0184	−0.0034	0.0010	0.0021	0.0655	−0.0062	0.0011	
	RMSE	0.0479	0.0498	0.0495	0.0411	0.0512	0.1233	0.0942	0.2365	-
BR without boundary obs. in the dataset	Bias	0.0150	0.0398	−0.0102	0.0013	0.0048	0.0937	−0.0207	−0.0454	
	RMSE	0.0499	0.0615	0.0507	0.0416	0.0515	0.1410	0.0995	0.2417	-
Zero-inflated beta regression	Bias	0.0152	0.0399	−0.0103	0.0013	0.0048	0.0934	−0.0208	−0.0454	0.0475
	RMSE	0.0500	0.0616	0.0507	0.0415	0.0515	0.1409	0.0995	0.2417	0.0517
Quasi-likelihood regression (p = 1)	Bias	−0.0080	−0.0047	−0.0008	−0.0003	0.0002	0.0932			
	RMSE	0.0583	0.0739	0.0635	0.0548	0.0658	0.1848	-	-	-

4.3 Third scenario

The third simulation scenario is similar to the first one in terms of there only being one explanatory variable generated by random draws from a standard uniform distribution. The covariate x was set for the sample size of n=50 and, equivalently to what was done in the first scenario, cloned to form the explanatory variable with n=100 observations where $x_i = x_{i+50}$, $i = 1, \dots, 50$, in that sample. The mean and precision are also given by the equations in (4.1) and (4.2), respectively. The difference in this scenario is that the data was generated according to a zero-inflated beta distribution as opposed to the regular beta distribution, which means that there is a probability α of each observation being equal to zero. Therefore, for every replicate of the simulation two samples were randomly generated (one with $n = 50$ and one with $n = 100$) such that

$$y_i \sim \text{BEZI}(\alpha, \mu_i, \phi_i), \;\; i = 1, \dots, n,$$

where μ_i and ϕ_i are obtained from applying the inverse logit function and the exponential function to (4.1) and (4.2), respectively. The parameter values were set as as $\beta_0 = 0.6$, $\beta_1 = 1.5$ and $\gamma = \log(60) = 4.094$, which results in values for the μ_i's that range from 0.372 to 0.710 with the median value being 0.538. Thus, the observations will tend to have values very far from the boundaries. The parameter α was set as 0.025, therefore on average 2.5% of the generated response value will be comprised of observations equal to zero that will be discrepant from the rest of the observations in the sample and will likely be highly influential on the model estimates. This scenario serves to analyse how the models discussed in this work fare in a hypothetical type of data where the boundary observations have a particular meaning and their nature is such that they are outliers. In practice, this could refer to a dataset where the zeros indicate an absence of something, however when this "something" is present, it accounts for a proportion of the total that is not too close to zero, an example of this is discussed in Section 2.2.1 with the proportion of monthly income spent with mobile phone services. This scenario is a little more extreme than this example as it is very unlikely that mobile services will cost near half of one's income, but the idea is to assess how much the models are affected by the boundary observations being discrepant and in practice these would have a particular meaning. Thus, this situation differs greatly from the previous scenarios which emulated a limit detection problem when establishing the response variable's values, so the boundary observations were not too different to the rest of the observations. Naturally, the zero-inflated beta regression model is highly suitable for a situation such as this seeing as the generated data is known to be from the zero-inflated beta distribution and the observations equal to zero are, by definition, separate from the observations in (0,1), however when met with a situation such as this one in practice, where there are but few boundary observations in the dataset, one might be reluctant to use that model and try one that produces estimates for μ_i and ϕ_i taking all observations into account. Needless to say, due to the fact that there will be discrepant observations in the generated samples, the beta regression model with the robust estimation method will have its performance analysed in this scenario. The models fitted for this scenario were:

(1) Using the linear transformation proposed by Smithson and Verkuilen (2006) and

then fitting a beta regression (BR) model via maximum likelihood and the robust approach (RobBR) proposed by Ribeiro and Ferrari (2020).

(2) Adding 0.01 to the boundary observations and then fitting a beta regression (BR) model via maximum likelihood and the robust approach (RobBR) proposed by Ribeiro and Ferrari (2020).

(3) Adding 0.001 to the boundary observations and then fitting a beta regression (BR) model via maximum likelihood and the robust approach (RobBR) proposed by Ribeiro and Ferrari (2020).

(4) Replacing boundary observations with the smallest observation observed in the (0,1) interval in the dataset and and then fitting a beta regression (BR) model via maximum likelihood and the robust approach (RobBR) proposed by Ribeiro and Ferrari (2020).

(5) Removing boundary observations and and then fitting a beta regression (BR) model via maximum likelihood and the robust approach (RobBR) proposed by Ribeiro and Ferrari (2020).

(6) Fitting a one-inflated-beta regression model.

(7) Fitting a quasi-likelihood model proposed by Bonat *et al.* (2019).

With α being small, there was still a distinct probabilty of there being no zeros in the generated sample. Figure 4.3 shows the frequency of the amount of zeros in each generated sample. Note that just under 3000 generated samples of size 50 had no zeros present in the dataset and not many replicates had more than 3 zeros, whereas in the generated samples with 100 observations, less than a 1000 replicates had no zeros in the sample. As was the case in the previous simulation scenarios, in the event of an absence of boundary figures in the sample, a beta regression model was fitted without prior transformations to the data and the zero-inflated beta regression model was also replaced by a BR model.

Table 4.3 shows the results for the fitted models. In this scenario, out of the 10000 replicates, the quasi-likelihood model while estimating a value for p failed in 8087 when the sample had size 100 and in 7977 when it had size 50, therefore the results for this model are not shown in the table. Even though the results of the BR model fitted after replacing the boundary observations with the smallest obs. in (0,1) and the BR model after excluding the boundary observations are shown in the table, they are not adequate methods to analyse data in this scenario. These transformations are too aggressive in situations where the boundary observations are outliers, as they either alter the zeros drastically or remove them altogether. The results are displayed in order to provide a comparison between them and the other models that either use less drastic transformations or do not require any.

The BR models that were fitted after using less aggressive ways of transforming the data (linear transformation and adding an ε to the boundary observations) resulted in a significant bias to the estimates of the parameters, especially when it comes to the estimate of γ and note how all these biases are negative, therefore they are much lower than they should be. This is due to how much the beta regression model fit is being influenced by the boundary observations, resulting in a very poor accuracy of the estimates. Usually

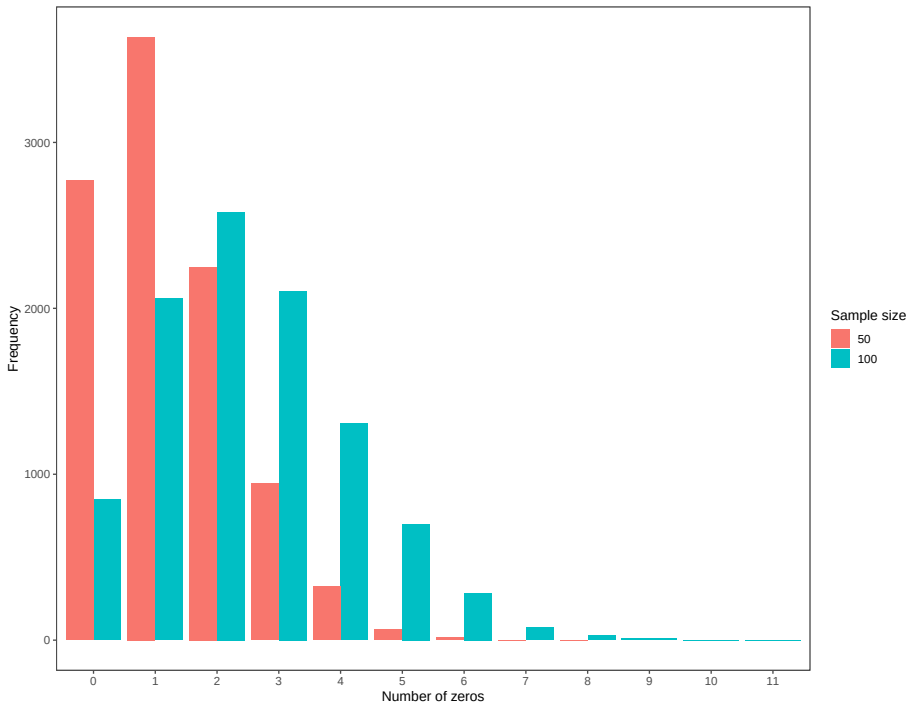


Figure 4.3: *Bar plot of the number of zeros in each generated sample for all 10000 replicates in the third simulation scenario.*

when the sample size increases, the accuracy of the estimates tend to increase, in this case however, as the presence of boundary observations implies that they are influent on the model estimates and a larger sample of 100 greatly decreases the amount of replicates in which the generated sample has no zeros and when the boundary observations are present, they tend to be more numerous than when they are present in the sample with size 50, the estimates are more affected in the larger sample.

The zero-inflated beta regression model naturally provides the best results as was expected given how the data was generated, but note how the robust beta regression model massively reduces the bias for the estimates of the parameters. In this scenario, the linear transformation provides a good alternative to deal with the boundary observations, provided a RobBr model is fitted, especially in the smaller sample where there are fewer observations equal to zero. In the larger sample size this method has worse results than adding an ε (0.01 or 0.001) to the boundary observations.

Note that in the smaller sample, the RobBR model fitted after adding 0.001 to the boundary observations provides better results than the RobBR after adding 0.01 to the zeros. This occurs due to 0.001 being more discrepant than 0.01 in relation to the observations in (0,1), henceforth the RobBR model attributes a lower weight to these observations than it does when they are equal to 0.01, therefore the robust estimation method is working as expected.

The QL model with p fixed as one seems to be less affected by the outliers than the BR models are, however the biases for the estimates is still very high, particularly when it comes to the estimation of γ.

In this scenario, aside from the zero-inflated beta regression model, the robust BR model fitted after using the linear transformation or after adding an ε to the boundary observations clearly provide the best alternatives for the analysis of these data, in both sample sizes. This type of method is able to deal well with discrepant observations and the transformations are not aggressive, therefore they do not alter the dataset too much and the model yields good results.

Table 4.3: *Estimated bias and root mean squared error (RMSE) for the estimates of the parameters in the third simulation scenario.*

		n =50				n =100			
Model		β_0	β_1	γ	α	β_0	β_1	γ	α
BR after using linear transformation	Bias	−0.035	−0.162	−1.368		−0.041	−0.181	−1.712	
	RMSE	0.181	0.395	1.673	-	0.147	0.339	1.848	-
BR after adding 0.01 to boundary obs.	Bias	−0.043	−0.142	−1.373		−0.041	−0.149	−1.585	
	RMSE	0.184	0.386	1.679	-	0.134	0.299	1.717	-
BR after adding 0.001 to boundary obs.	Bias	−0.060	−0.196	−1.669		−0.057	−0.209	−1.954	
	RMSE	0.247	0.515	2.019	-	0.182	0.406	2.096	-
BR after replacing boundary obs. with lowest obs. in (0,1)	Bias	−0.002	−0.048	−0.271		−0.003	−0.050	−0.358	
	RMSE	0.082	0.160	0.458	-	0.059	0.1221	0.459	-
BR without boundary obs. in the dataset	Bias	0.000	0.000	0.061		0.000	0.001	0.030	
	RMSE	0.075	0.125	0.216	-	0.053	0.088	0.148	-
Zero-inflated beta regression	Bias	0.000	0.000	0.061	0.000	0.000	0.000	0.030	0.000
	RMSE	0.075	0.125	0.216	0.022	0.053	0.088	0.148	0.016
Quasi-likelihood regression (p = 1)	Bias	−0.038	−0.043	0.868		−0.035	−0.043	−0.977	
	RMSE	0.119	0.217	1.110	-	0.086	0.157	1.096	-
RobBR after using linear transformation	Bias	0.006	−0.036	0.001		0.005	−0.016	0.021	
	RMSE	0.091	0.181	0.506	-	0.055	0.096	0.226	-
RobBR after adding 0.01 to boundary obs.	Bias	−0.004	−0.012	−0.032		−0.001	0.000	−0.001	
	RMSE	0.092	0.180	0.505	-	0.054	0.093	0.203	-
RobBR after adding 0.001 to boundary obs.	Bias	−0.003	−0.010	−0.011		0.000	−0.002	−0.007	
	RMSE	0.111	0.216	0.482	-	0.056	0.105	0.235	-
RobBR after replacing boundary obs. with lowest obs. in (0,1)	Bias	−0.013	−0.006	−0.113		−0.014	−0.004	−0.173	
	RMSE	0.082	0.144	0.329	-	0.059	0.100	0.281	-
RobBR without boundary obs. in the dataset	Bias	0.000	0.000	0.061		0.000	0.00	0.030	
	RMSE	0.075	0.125	0.216	-	0.053	0.088	0.148	-

Chapter 5

Conclusions

In this work, the performance of different methods to deal with boundary observations was compared in different scenarios. In the wealth of situations analysed, it is possible to come to some conclusions as to the performance of the methods and how much sense each of them makes in each scenario.

When discussing methods to map data onto the (0,1) interval, two of the methods in particular are, in general, very aggressive: replacing the boundary observations with the value of the smallest/largest observation in (0,1) and excluding the boundary observations. Even though these are, theoretically, viable options in order to deal with the obstacle of having boundary observations, they usually do not make much sense. Excluding some observations is often not advisable as they can still provide useful information about the data and what is being studied, unless one possesses the knowledge that these observations being analysed have been wrongly measured or if there is another problem with it. For instance, in the tuna application the sole boundary observation is completely discrepant from all other observations, but unless it is known for certain that the observation is wrong or does not make sense, excluding it would be discarding a possible outcome of the data. However this transformation is useful as a way to compare what the model would be like in the absence of boundary observations and since this work is mainly about cases with few of these, it is desirable that the estimates are not too altered when fitting a model with the full data. Replacing by the smallest/largest observation in (0,1) equates the boundary observation to another which may have completely different characteristics. One could argue that in situations where the values are all small (such as in the CVE application), the boundary observations are very close to the observations with smallest values in (0,1), however the difference relative to the magnitude of the data may not be small at all. Equating the boundary observations to observations that may be close is not adequate, seeing as it is interesting that any transformations used maintain a difference from observations equal to 0 or to 1 to the rest. In situations where the boundary observations are discrepant, this method would completely alter the data and is not an adequate way of dealing with boundaries.

When it comes to mapping the data onto (0,1), the most used method in works published in recent years has been to use the linear transformation from Smithson and Verkuilen (2006). However, as seen in some scenarios explored in this work, this transformation is not

always a good alternative. As seen in the CVE application, when the response variable's values are low, the transformation alters the data to a point where the model fitted after using it yields poor estimates for the parameters. This situation is confirmed in the second scenario in Chapter 4, as even with a generated sample, the transformation continues to provide worse results than other alternative ways to deal with boundary observations. Therefore, this method cannot be used without taking the context of the data into account; depending on the characteristics of the response variable, this transformation may greatly affect the model estimates, consequently affecting their interpretation. Henceforth, caution is advised when deciding if this transformation should be used.

Adding or subtracting an ε to the boundary observations was shown to be a good alternative in a variety of scenarios. Choosing an appropriate ε allows the boundary observations to still be different from the rest of the observations, unlike when using the more aggressive method already discussed, and maintains the structure of the data, which is not always the case when using the linear transformation. Despite being a degree of subjectivity when choosing which ε is more appropriate for each scenario, a brief analysis of the data should be enough to better decide on which ε to use and more than one option can be tested to ensure a good choice. In cases where the boundary observations are discrepant, the method of adding (subtracting) an ε worked really well alongside the beta regression model with the robust estimation method. With this method, the observations will still be very close to the boundary while the RobBR attributes a smaller weight to these observations when estimating the parameters. The robust estimation method for the beta regression model is clearly the best way to deal with discrepant boundary observations if the use of the inflated beta regression model is not adequate. The third simulation scenario shows that this method is the option that is least affected by the outliers equal to zero or one and as previously stated, adding (subtracting) a ε to the boundaries and then fitting a BR model with this robust estimation method works well. In fact, in many cases whatever the choice of a small ε the estimates for the beta regression model with the robust approach may not be altered too much at all.

The advantage of opting for the traditional quasi-likelihood approach ($p = 1$) is that it is less affected by the boundary observations than the BR model and no prior transformation of the data is required. However, in situations where the boundaries are highly influent on the model, the RobBR is without a doubt the best choice. If the boundaries are not influent, the scenarios studied in this paper showed that the QL approach is not as good as the BR model, except in situations where the assumption of a fixed precision is reasonable, as seen in the first simulation scenario, therefore it may still be useful in such situations. However, in practice, the assumption of a fixed precision is rarely appropriate in the analysis of data in the (0,1) interval. When estimating p, the influent boundary observations greatly affect the estimation of p itself, thereby causing the estimated dispersion to be completely distorted to the point where its interpretation will no longer make much sense. Also, due to the limitation of the implementations of the QL models being restricted to a fixed precision (or dispersion), in many situation the QL model will not be able to accurately estimate the precision parameter, which was the case in the CVE application and further explored in the simulation scenarios. Even without influent boundary observations, the QL model while estimating p failed to work in many simulated situations, which is not ideal given that theoretically having a non-fixed p would increase the flexibility of the model and

assist in it providing decent models in more data scenarios. However, the presence of the non-fixed p seemed to limit the model even more, at least in situations with boundary observations, henceforth being unreliable. Another disadvantage of opting for the QL approach is the lack of diagnostic tools normally available.

The inflated beta regression model is naturally more adept to scenarios where there is a larger cluster at the boundaries than in the scenarios explored here, however it is important to note that in situations like the CVE application where there is a reasonable amount of observations and the boundaries account to 5% of them, this model is a viable option despite the occurrence of a zero in the data not having a special meaning. Therefore, this type of model should still be considered and it may be useful in some scenarios, but one must be careful with the diagnostics; as seen in Section 3.2, the randomized quantile residual for the inflated beta regression models is not correctly programmed in the `gamlss` package on `R`, thus it it advisable to calculate the residual separately in order to avoid an interpretation that the model is not well fitted when it in fact may be.

In summary, adding (subtracting) ε to (from) the boundaries seems to be a method that is adequate in more scenarios, but the linear transformation may also be an option in some cases. In situations where the boundaries are influent, the robust beta regression model is an excellent choice and the inflated beta regression may be a viable option depending on the amount of boundary observations in the sample.

Essentially, there are no methods that can be used indiscriminately without prior analysis. The most adequate method needs to be carefully analysed and some situations that boundary observations may arise were discussed in this thesis. Even though the linear transformation is often used to transform the data, some examples presented here show that it is not always a good alternative, just like a poor choice of ε to be added (subtracted) to (from) the boundary observation may result in inaccurate estimates for the parameters.

Appendix A

Diagnostics - Tuna application

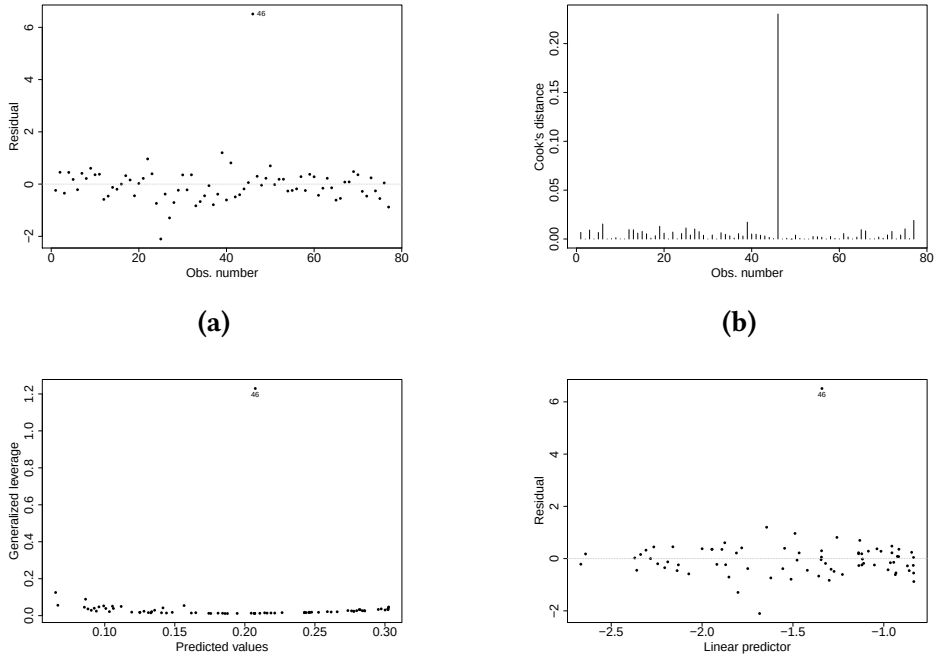


Figure A.1: *Diagnostic graphs for the BR model fitted after using the linear transformation in the tuna application.*

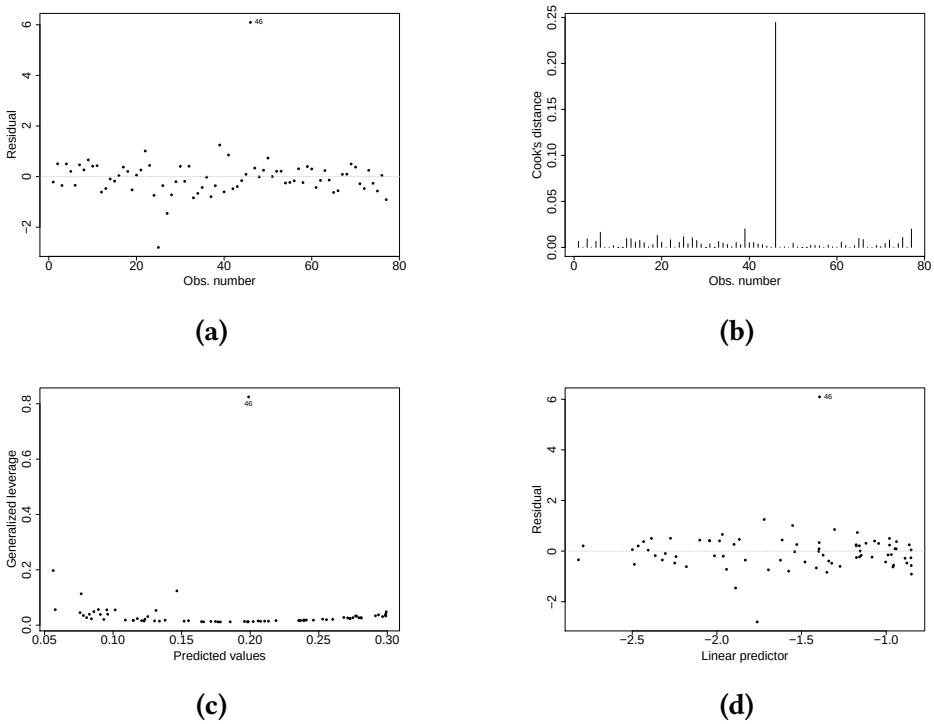


Figure A.2: *Diagnostic plots for the BR model fitted after subtracting 0.01 from the boundary observation in the tuna application.*

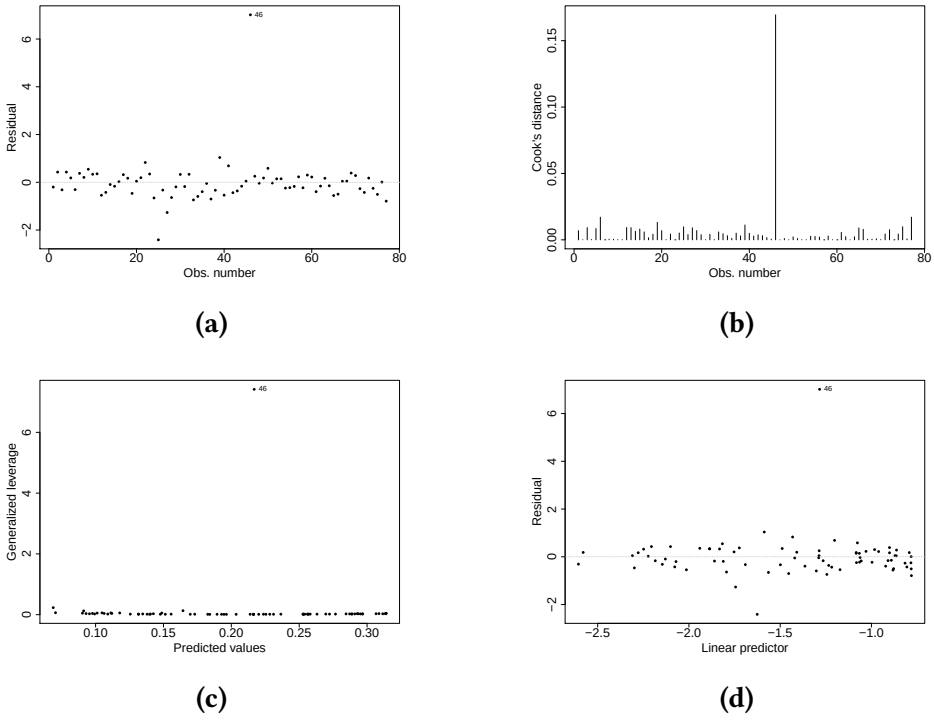


Figure A.3: *Diagnostic plots for the BR model fitted after subtracting 0.001 from the boundary observation in the tuna application.*

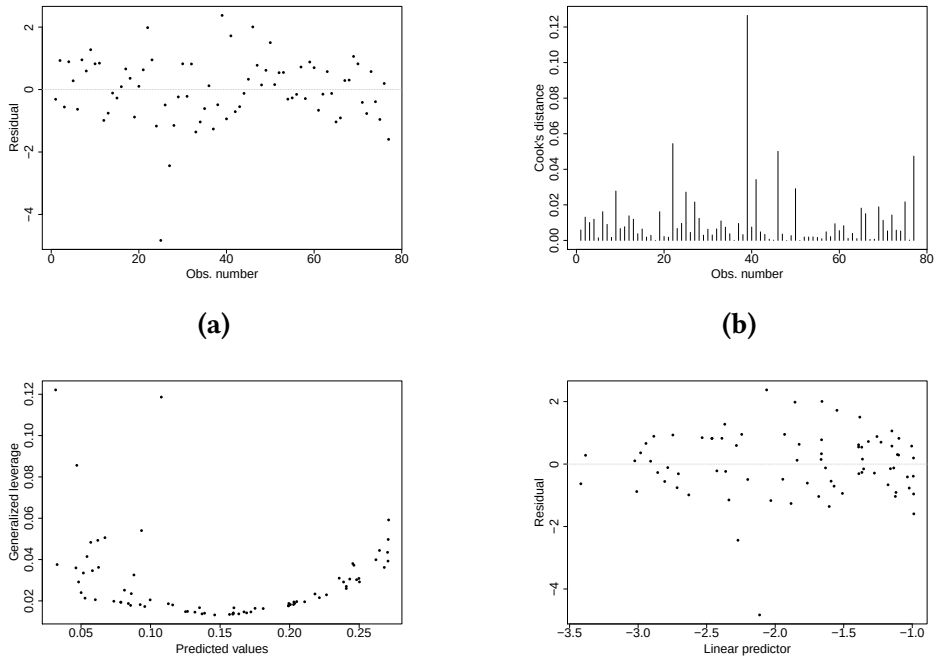


Figure A.4: *Diagnostic plots for the BR model fitted after replacing boundary observations with largest obs. in (0,1) in the tuna application.*

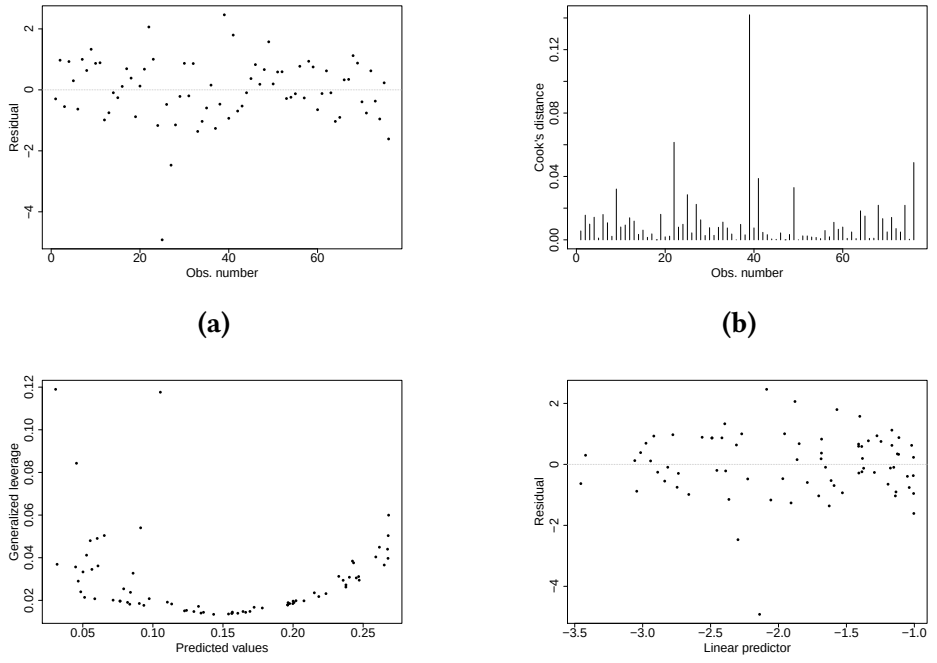
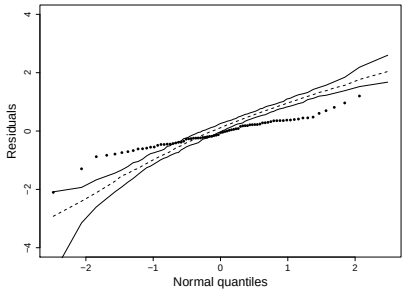
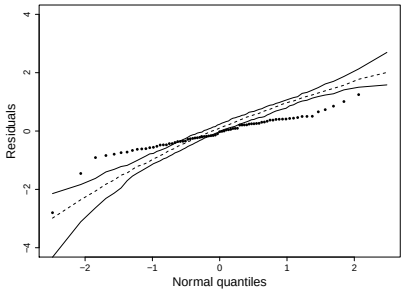


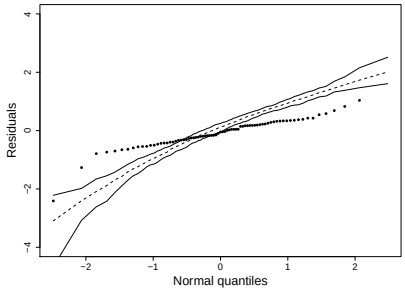
Figure A.5: *Diagnostic plots for the BR model fitted after removing boundary observations in the tuna application.*



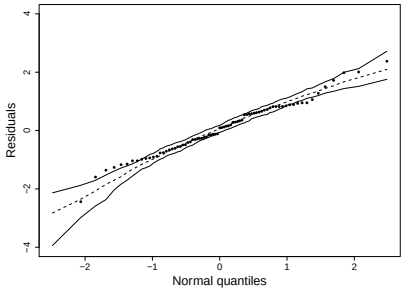
(a) *BR after using linear transformation*



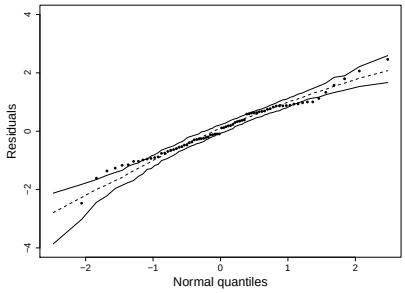
(b) *BR after subtracting 0.01 from obs. 46*



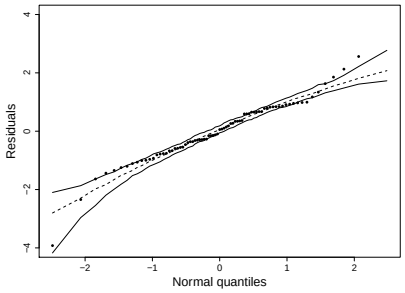
(c) *BR after subtracting 0.001 from obs. 46*



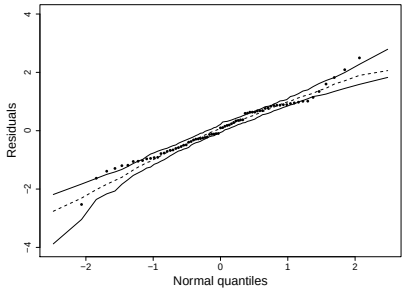
(d) *BR after replacing obs. 46 with largest obs. in (0,1)*



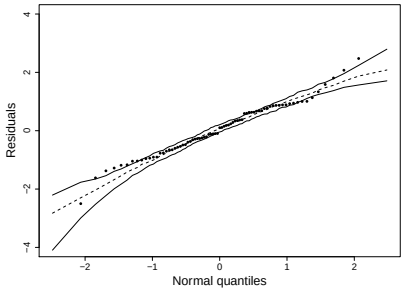
(e) *BR after removing obs. 46*



(f) *RobBR after using linear transformation*



(g) *RobBR after subtracting 0.01 from obs. 46*



(h) *RobBR after subtracting 0.001 from obs. 46*

Figure A.6: *Normal probability plots with simulated envelopes for the models fitted in the tuna application, where the residuals used are the standardized weighted residual type 2.*

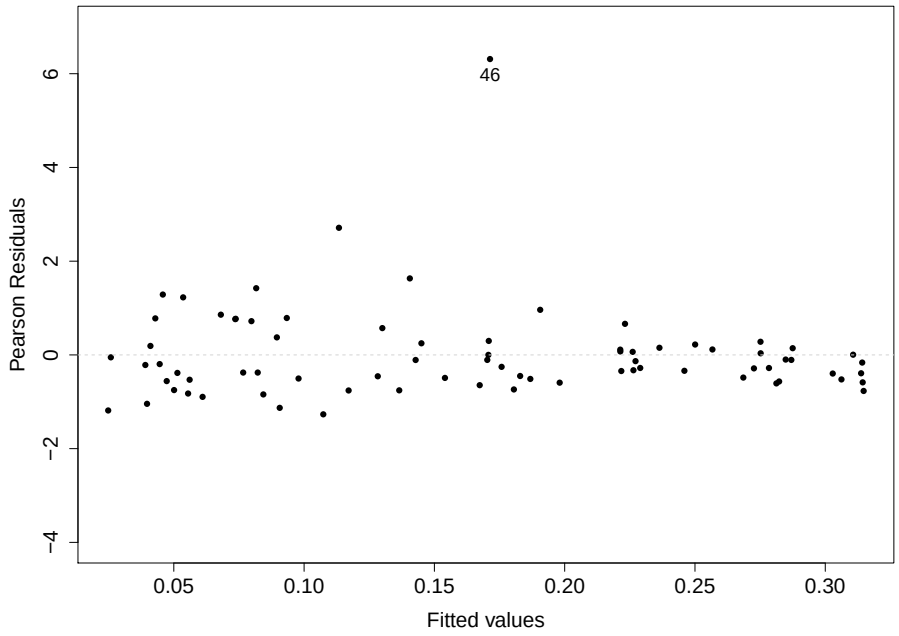


Figure A.7: *Scatter plot of the Pearson residuals vs. the fitted values of the quasi-likelihood model with an estimated p fitted for the analysis of the tuna application.*

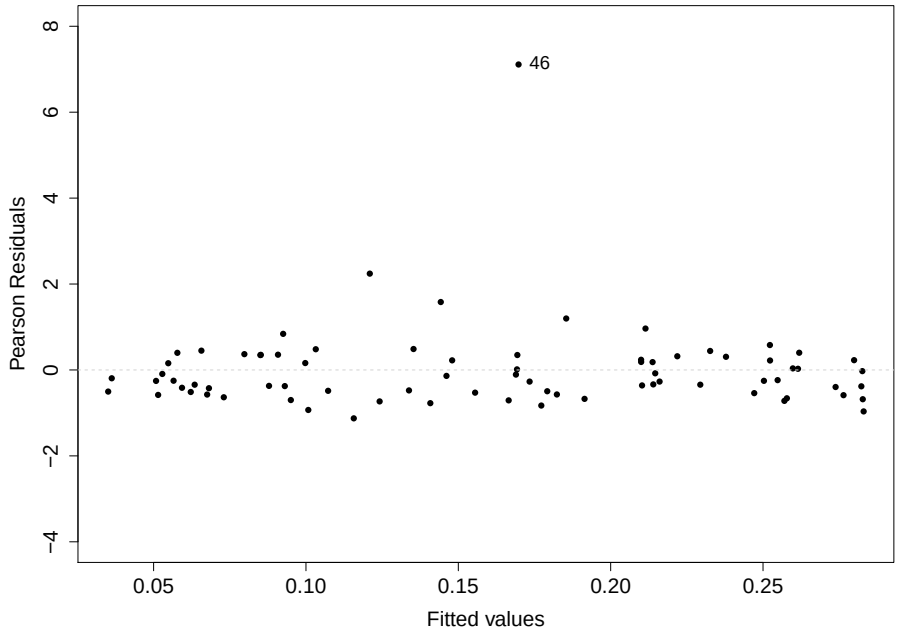
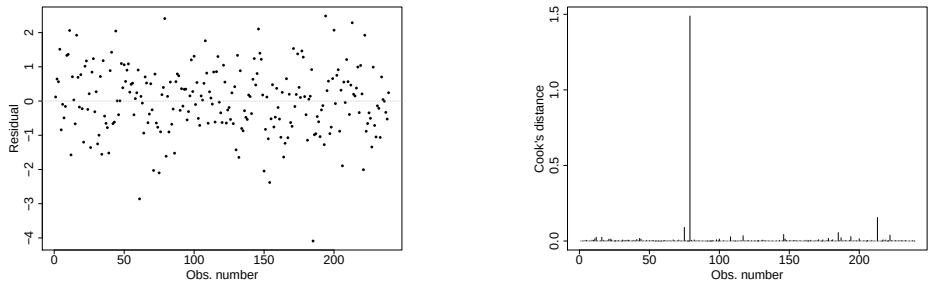


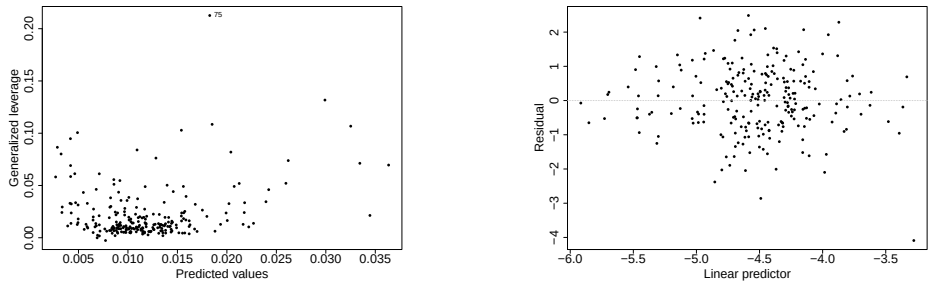
Figure A.8: *Scatter plot of the Pearson residuals vs. the fitted values of the quasi-likelihood model with p=1 fitted for the analysis of the tuna application.*

Appendix B

Diagnostics - CVE application



(a) **(b)**



(c) **(d)**

Figure B.1: *Diagnostic graphs for the BR model fitted after using the linear transformation in the CVE application.*

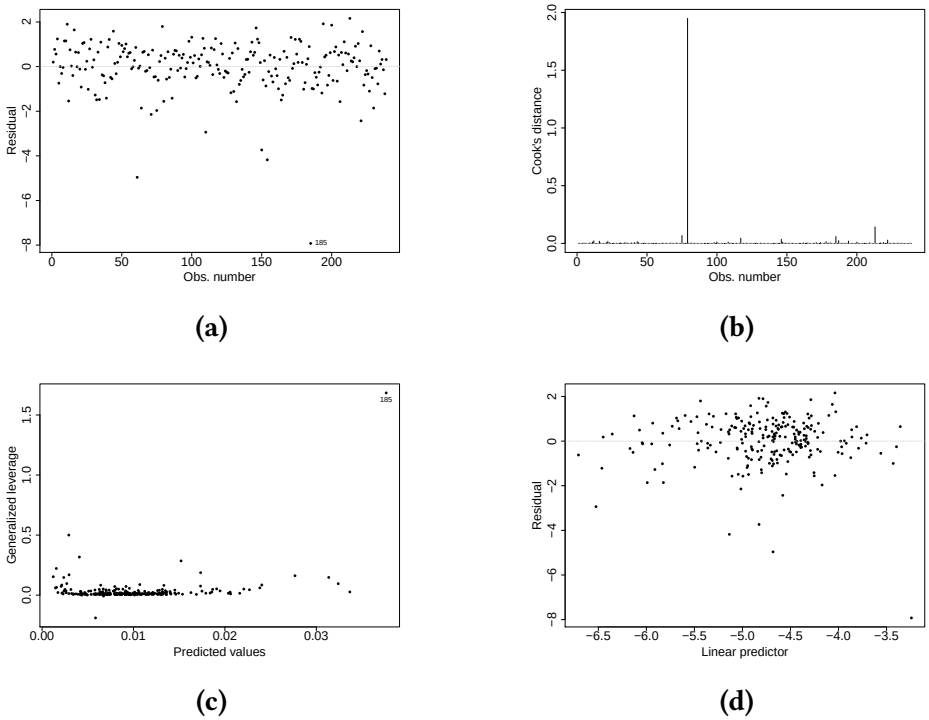


Figure B.2: *Diagnostic graphs for the BR model fitted after adding 0.0001 to the boundary observations in the CVE application.*

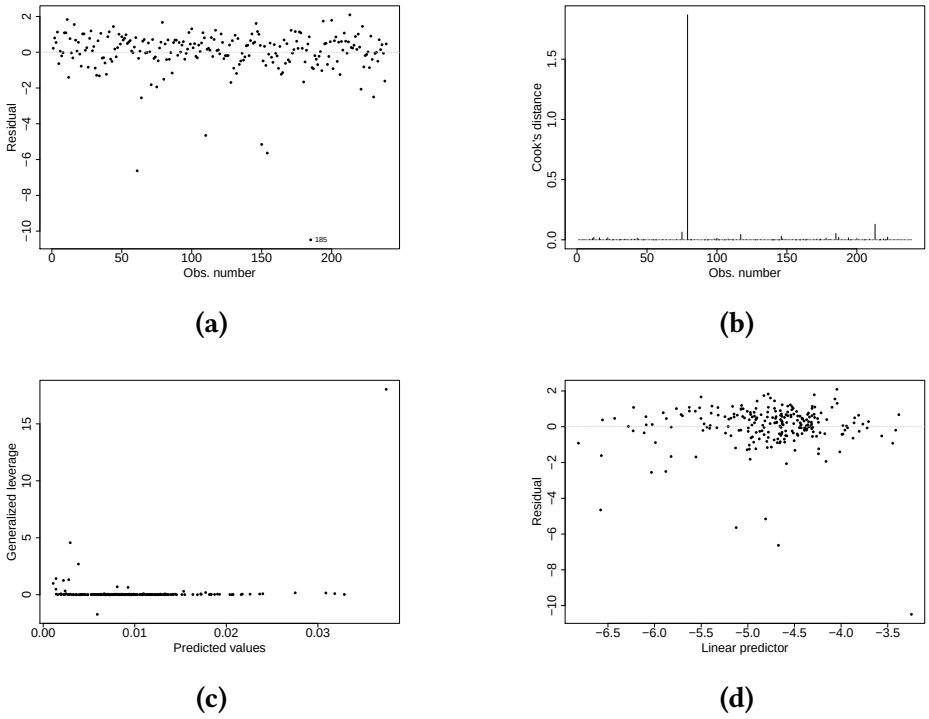


Figure B.3: *Diagnostic plots for the BR model fitted after adding 0.00001 to the boundary observations in the CVE application.*

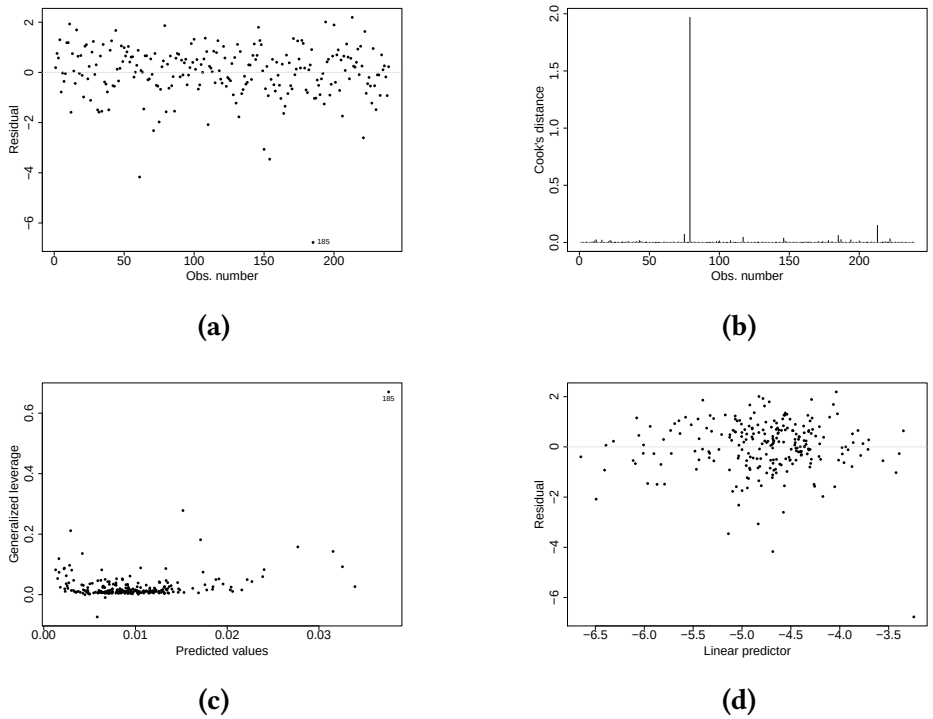


Figure B.4: *Diagnostic plots for the BR model fitted after replacing boundary observations with smallest in (0,1) in the CVE application.*

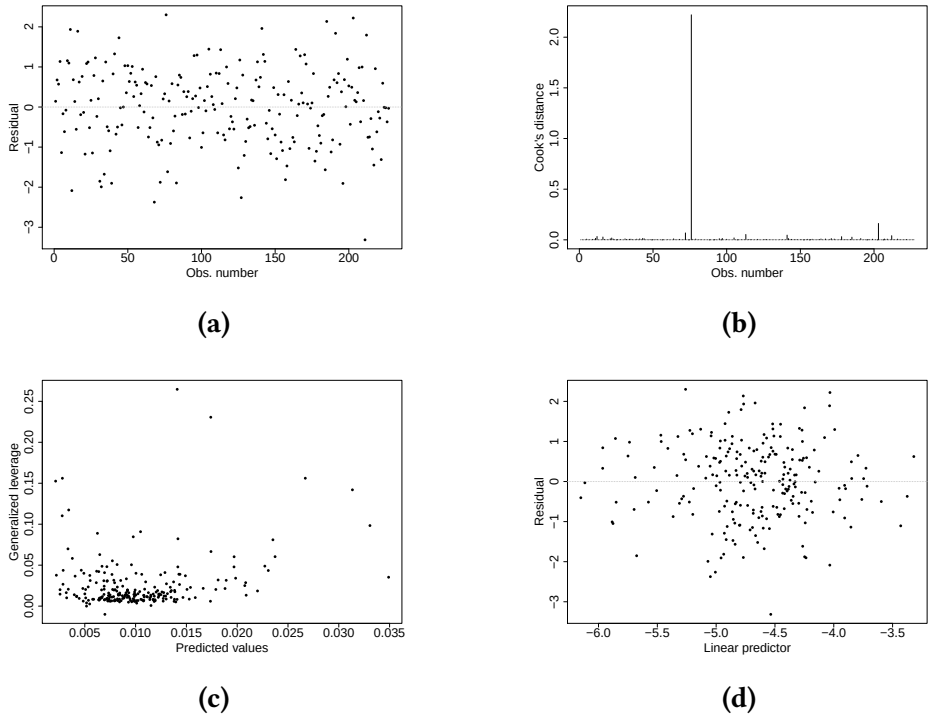
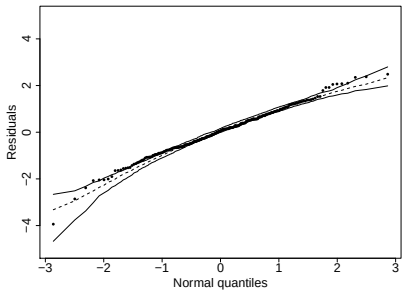
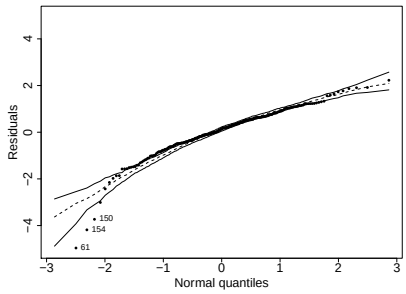


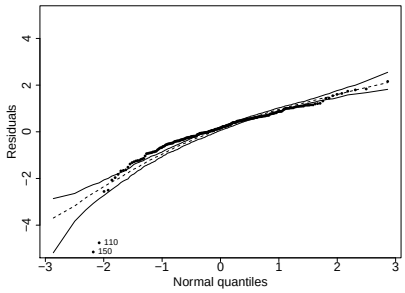
Figure B.5: *Diagnostic plots for the BR model fitted after removing boundary observations in the CVE application.*



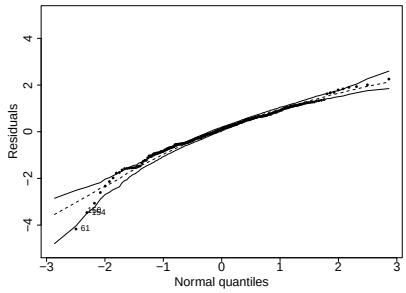
(a) *BR after using linear transformation*



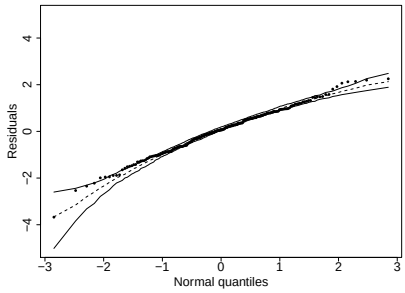
(b) *BR after subtracting 0.0001 from obs. 46*



(c) *BR after subtracting 0.00001 from obs. 46*



(d) *BR after replacing obs. 46 with smallest obs. in (0,1)*



(e) *BR after excluding boundary observations.*

Figure B.6: *Normal probability plots with simulates envelopes for the five beta regression models fitted via maximum likelihood in the CVE application. Each figure is captioned according to what method was fitted.*

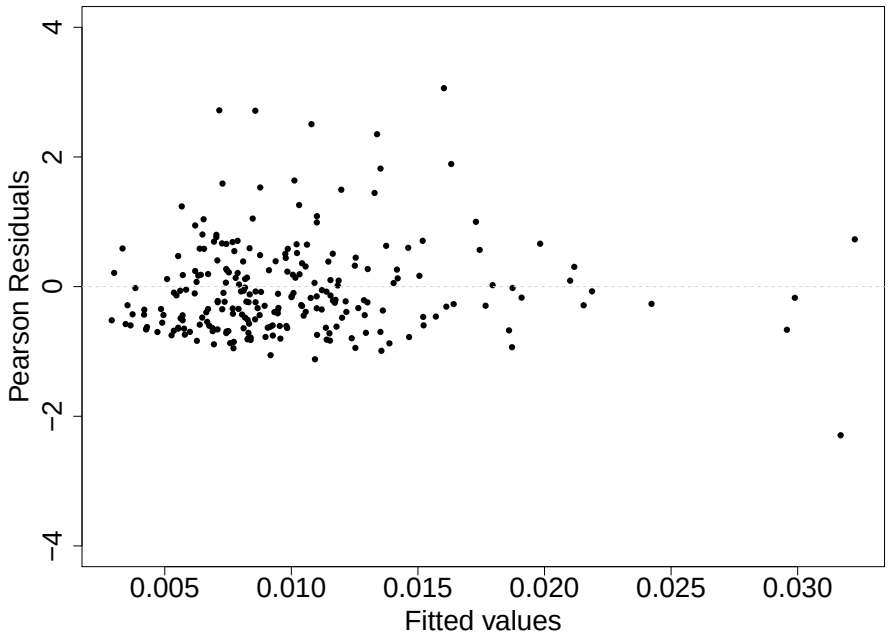


Figure B.7: *Scatter plot of the Pearson residuals vs. the fitted values of the quasi-likelihood model fitted in the CVE application.*

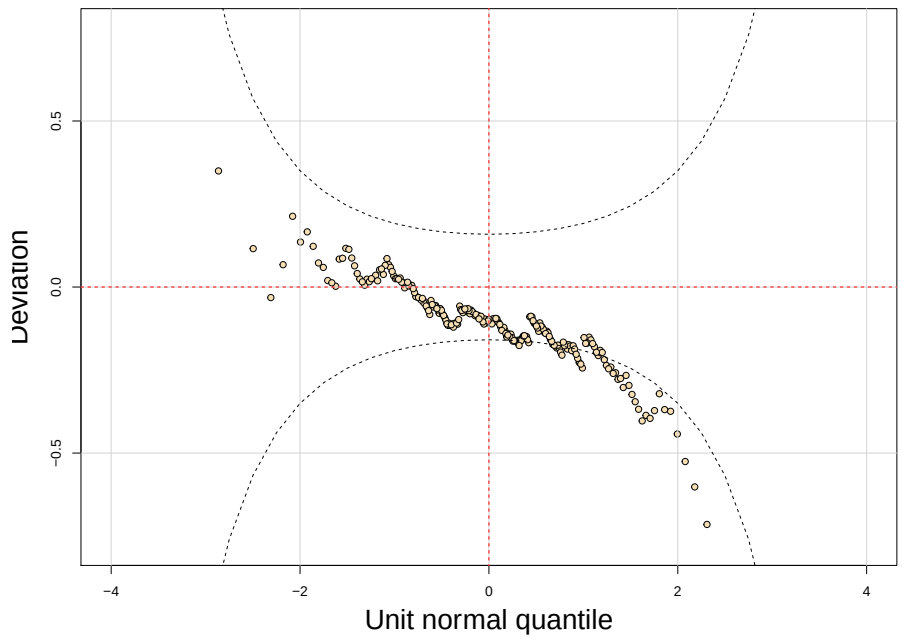


Figure B.8: *Worm plot for the inflated beta regression model fitted for the CVE data using the residual calculated with the* `gamlss` *package on* `R`.

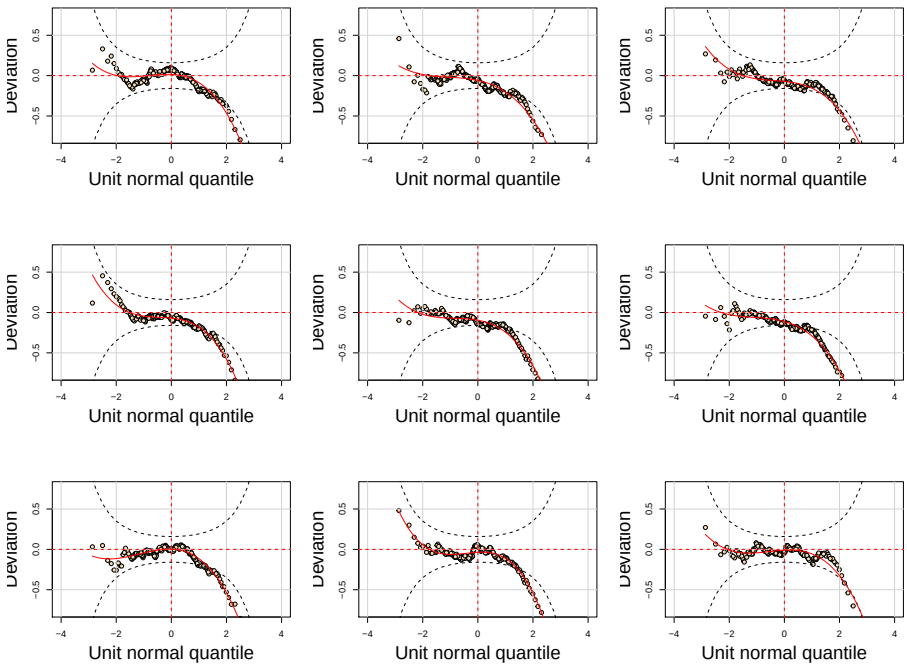


Figure B.9: *Worm plots of the fitted models for the nine generated samples.*

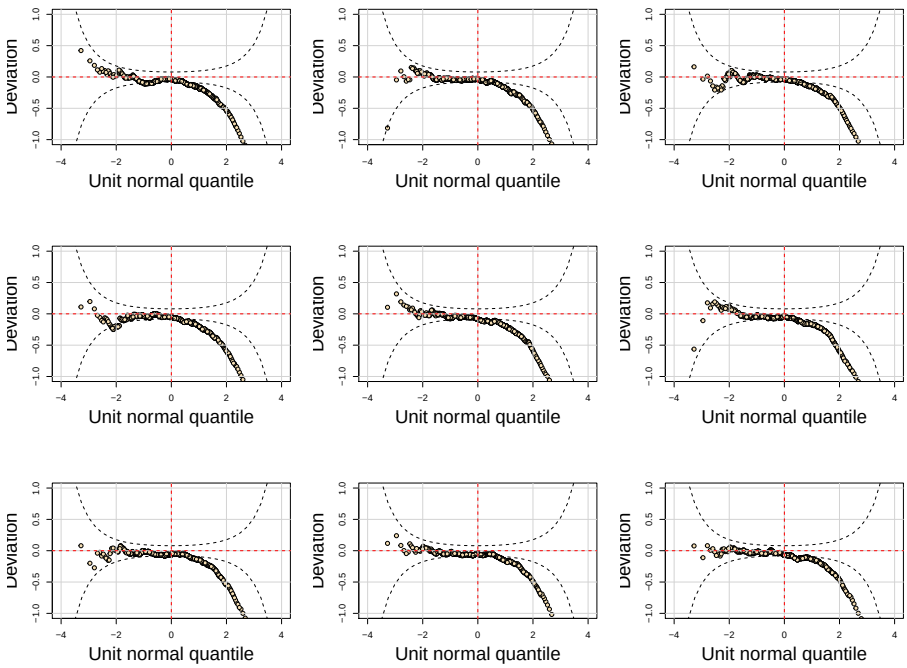


Figure B.10: *Worm plots of the fitted models for the nine generated samples with the quadruple sample size.*

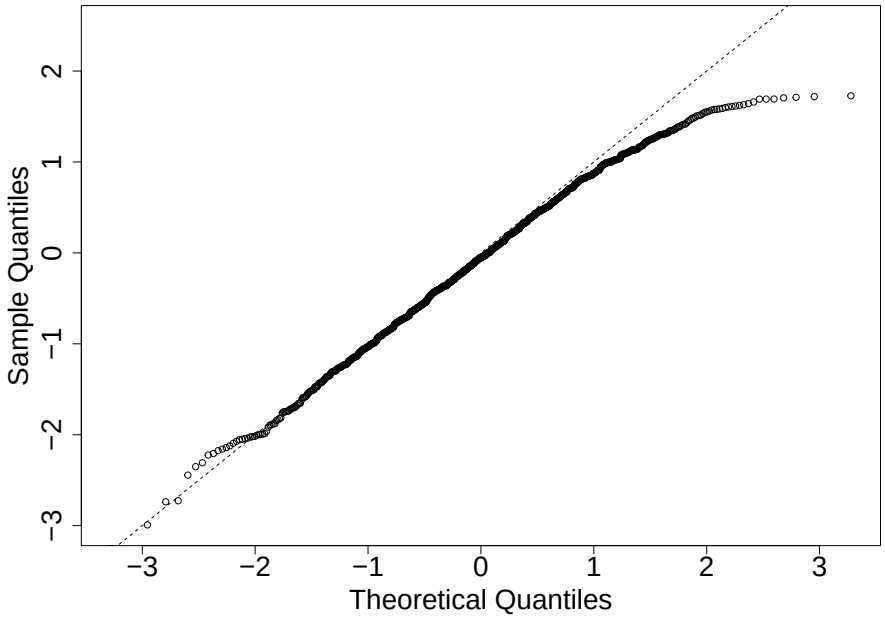


Figure B.11: *Normal probability plot of a generated data with quadruple sample size, where the randomized quantile residuals were used in comparison to the normal quantiles.*

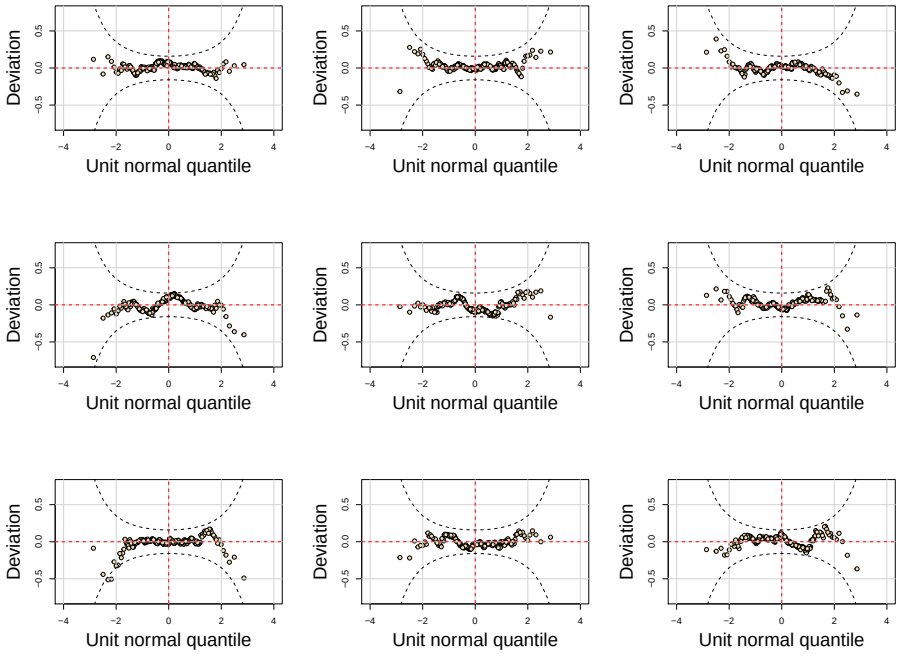


Figure B.12: *Worm plots of the fitted models for the nine generated samples using the correct expression for the r_q residual.*

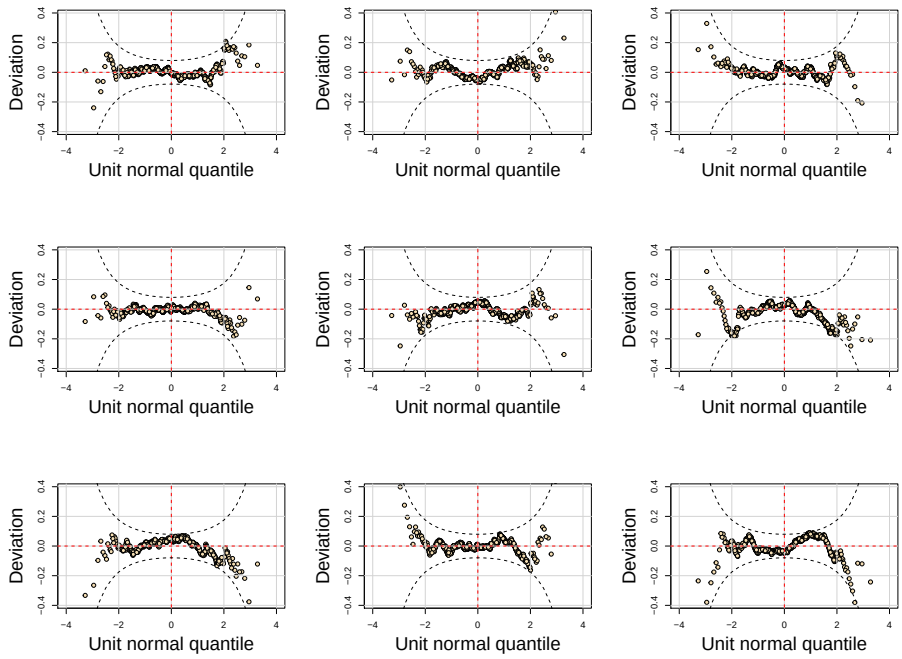


Figure B.13: *Worm plots of the fitted models for the nine generated samples with the quadruple sample size using the correct expression for the r_q residual.*

References

[Bayes *et al.* 2012] Cristian Bayes, Jorge Bazán, and Catalina Garcia. "A New Robust Regression Model for Proportions". *Bayesian Analysis* 7 (Aug. 2012), pp. 841–866. doi: 10.1214/12-BA728 (cit. on p. 9).

[Bonat and Jørgensen 2015] Wagner Bonat and Bent Jørgensen. "Multivariate Covariance Generalized Linear Models". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* (Apr. 2015), pp. 649–675. doi: 10.1111/rssc.12145 (cit. on p. 10).

[Bonat *et al.* 2019] Wagner Bonat, Ricardo R. Petterle, John Hinde, and Clarice G. B. Demétrio. "Flexible quasi-beta regression models for continuous bounded data". *Statistical Modelling* 19.6 (2019), pp. 617–633. doi: 10.1177/1471082X18790847 (cit. on pp. 9, 11, 15, 25, 34, 38, 43).

[Buuren and Fredriks 2001] Stef Buuren and Miranda Fredriks. "Worm plot: A simple diagnostic device for modelling growth reference curves". *Statistics in Medicine* 20 (Apr. 2001), pp. 1259–77. doi: 10.1002/sim.746 (cit. on p. 13).

[Di Brisco *et al.* 2020] Agnese Maria Di Brisco, Sonia Migliorati, and Andrea Ongaro. "Robustness against outliers: A new variance inflated regression model for proportions". *Statistical Modelling* 20.3 (2020), pp. 274–309. doi: 10.1177/1471082X18821213 (cit. on p. 9).

[Douma and Weedon 2019] Jacob Douma and James Weedon. "Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression". *Methods in Ecology and Evolution* 10 (June 2019). doi: 10.1111/2041-210X.13234 (cit. on p. 7).

[Dunn and Smyth 1997] Peter Dunn and Gordon Smyth. "Randomized Quantile Residuals". *Journal of Computational and Graphical Statistics* 5 (Aug. 1997). doi: 10.1080/10618600.1996.10474708 (cit. on p. 14).

[Espinheira *et al.* 2008] Patrícia L. Espinheira, Silvia L. P. Ferrari, and Francisco Cribari-Neto. "On beta regression residuals". *Journal of Applied Statistics* 35.4 (2008), pp. 407–419. doi: 10.1080/02664760701834931 (cit. on p. 14).

[Ferrari and Cribari-Neto 2004] Silvia L. P. Ferrari and Francisco Cribari-Neto. "Beta Regression for Modelling Rates and Proportions". *Journal of Applied Statistics* 31.7 (2004), pp. 799–815. URL: https://EconPapers.repec.org/RePEc:taf:japsta:v:31:y:2004:i:7:p:799-815 (cit. on pp. 1, 3, 4).

[Hunger *et al.* 2012] Matthias Hunger, Angela Döring, and Rolf Holle. "Longitudinal beta regression models for analyzing health-related quality of life scores over time". *BMC Medical Research Methodology* 12 (Sept. 2012), p. 144. DOI: 10.1186/1471-2288-12-144 (cit. on pp. 5, 6).

[Jørgensen 1997] Bent Jørgensen. "Proper Dispersion Models". *Brazilian Journal of Probability and Statistics* 11.2 (1997), pp. 89–128. ISSN: 01030752, 23176199. URL: http://www.jstor.org/stable/43600934 (cit. on p. 1).

[Jørgensen and Knudsen 2004] Bent Jørgensen and Sven J. Knudsen. "Parameter Orthogonality and Bias Adjustment for Estimating Functions". *Scandinavian Journal of Statistics* 31.1 (2004), pp. 93–114. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9469.2004.00375.x (cit. on p. 10).

[Kieschnick and Mccullough 2003] Robert Kieschnick and B. D. Mccullough. "Regression analysis of variates observed on (0 1)". *Statistical Modelling - STAT MODEL* 3 (Oct. 2003), pp. 193–213. DOI: 10.1191/1471082X03st053oa (cit. on p. 9).

[Lemonte and Bazán 2016] Artur J. Lemonte and Jorge L. Bazán. "New class of Johnson distributions and its associated regression model for rates and proportions". *Biometrical Journal* 58.4 (2016), pp. 727–746. DOI: 10.1002/bimj.201500030 (cit. on p. 9).

[Lima-Filho *et al.* 2020] Luiz M. A. Lima-Filho, Tarciana Liberal Pereira, Tatiene C. Souza, and Fábio M. Bayer. "Process monitoring using inflated beta regression control chart". *PLOS ONE* 15.7 (July 2020), pp. 1–20. DOI: 10.1371/journal.pone.0236756 (cit. on p. 5).

[Malik and Piepho 2016] Waqas Ahmed Malik and Hans-Peter Piepho. "On generalized exponential transformations for proportions". *Communications in Statistics - Theory and Methods* 45.19 (2016), pp. 5857–5870. DOI: 10.1080/03610926.2014.950753 (cit. on p. 2).

[Migliorati *et al.* 2017] Sonia Migliorati, Agnese Maria Di Brisco, and Andrea Ongaro. "A New Regression Model for Bounded Responses". *Bayesian Analysis* 13 (Oct. 2017), pp. 845–872. DOI: 10.1214/17-BA1079 (cit. on p. 9).

[Mitnik and Baek 2013] Pablo A. Mitnik and Sunyoung Baek. "The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation". *Statistical Papers* 54.1 (2013), pp. 177–192. DOI: 10.1007/s00362-011-0417-y (cit. on pp. 1, 9).

[Monllor-Hurtado *et al.* 2017] Alberto Monllor-Hurtado, Maria Pennino, and José Sánchez Lizaso. "Shift in tuna catches due to ocean warming". *PLOS ONE* 12 (June 2017). doi: 10.1371/journal.pone.0178196 (cit. on p. 14).

[Morrison *et al.* 2020] Maike Morrison, Lauren Castro, and Lauren Meyers. "Conscientious vaccination exemptions in kindergarten to eighth-grade children across Texas schools from 2012 to 2018: A regression analysis". *PLOS Medicine* 17 (Mar. 2020), e1003049. doi: 10.1371/journal.pmed.1003049 (cit. on pp. 6, 23, 26, 30).

[Ospina and Ferrari 2010] Raydonal Ospina and Silvia L. P. Ferrari. "Inflated beta distributions". *Statistical Papers* 51.1 (Mar. 2010), pp. 111–126. issn: 1613-9798. doi: 10.1007/s00362-008-0125-4 (cit. on pp. 7, 9).

[Ospina and Ferrari 2012] Raydonal Ospina and Silvia L. P. Ferrari. "A general class of zero-or-one inflated beta regression models". *Computational Statistics & Data Analysis* 56.6 (2012), pp. 1609–1623. issn: 0167-9473. doi: https://doi.org/10.1016/j.csda.2011.10.005 (cit. on pp. 2, 9).

[Pereira 2019] Gustavo Pereira. "On quantile residuals in beta regression". *Communications in Statistics - Simulation and Computation* 48 (2019), pp. 302–316. doi: 10.1080/03610918.2017.1381740 (cit. on p. 14).

[Piepho 2003] Hans-Peter Piepho. "The folded exponential transformation for proportions". *Journal of the Royal Statistical Society: Series D (The Statistician)* 52.4 (2003), pp. 575–589. doi: https://doi.org/10.1046/j.0039-0526.2003.00509.x (cit. on p. 2).

[R Core Team 2020] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020. url: https://www.R-project.org/ (cit. on pp. 2, 11, 13).

[Ribeiro and Ferrari 2020] Terezinha K. A. Ribeiro and Silvia L. P. Ferrari. "Robust estimation in beta regression via maximum Lq-likelihood" (Oct. 2020) (cit. on pp. 4, 5, 13–15, 18, 43).

[Rudolf *et al.* 2019] J. Rudolf, Lu-Yi Wang, Stanislav Gorb, and Hamed Rajabi. "On the fracture resistance of dragonfly wings". *Journal of the Mechanical Behavior of Biomedical Materials* 99 (July 2019). doi: 10.1016/j.jmbbm.2019.07.009 (cit. on pp. 1, 2, 5).

[Schmid *et al.* 2013] Matthias Schmid *et al.* "Boosted Beta Regression". *PLOS ONE* 8 (Apr. 2013), e61623. doi: 10.1371/journal.pone.0061623 (cit. on p. 6).

[Smithson and Verkuilen 2006] Michael Smithson and Jay Verkuilen. "A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables". *Psychological Methods* 11 (Apr. 2006), pp. 54–71. doi: 10.1037/1082-989X.11.1.54 (cit. on pp. 1, 6, 14, 17, 25, 34, 38, 42, 47).