FACULTAD DE INGENIERÍA
REVISTA DE LA FACULTAD DE INGENIERIA
www.revistaingenieria.uda.cl

Beta Model to Investigate the Proportion of Deaths Caused by Acute Diarrheal Disease

Oilson Alberto Gonzatto Junior[1], Hellen Geremias dos Santos[2,3], and Marcos Jardel Henriques[3]

[1]Institute of Mathematical and Computer Sciences, University of São Paulo (ICMC-USP), São Carlos, São Paulo, Brazil.
[2]Oswaldo Cruz Foundation (FIOCRUZ). Curitiba, Paraná, Brazil.
[3]Interinstitutional Postgraduate Program in Statistics (PIPGEs) UFSCar-USP (Federal University of São Carlos (DES-UFSCar) and University of São Paulo (ICMC-USP)). São Carlos, São Paulo, Brazil. Email: jardel@usp.br

ABSTRACT

Acute diarrheal disease is a syndrome caused by different etiological agents. In the 1980s, the disease caused over 17% of infant deaths. Government measures to control the disease were significant, and the latest report from the Ministry of Health in 2011 indicated a decrease in infant deaths caused by this disease to 1.6%. This study used records provided by the Ministry of Health for the proportions of deaths due to acute diarrheal disease in children under 5 years old in each federation unit from 1990 to 2011. These proportions were adjusted using a beta regression model, with region and year of occurrence as covariates. The analysis performed on the model corroborated clinical findings highlighted in the literature, which demonstrate the effectiveness of the measures taken over the years to combat this disease.

Keywords: Infant Mortality, Beta Regression, Acute Diarrheal Disease, Clinical Findings, Disease Control

RESUMEN

La enfermedad diarreica aguda es un síndrome causado por diferentes agentes etiológicos. En la década de los 80, la enfermedad causó más del 17% de las muertes infantiles. Las medidas gubernamentales para controlar la enfermedad fueron significativas, y el último informe del Ministerio de Salud en 2011 indicó una disminución de las muertes infantiles causadas por esta enfermedad al 1.6%. Este estudio utilizó los registros proporcionados por el Ministerio de Salud para las proporciones de muertes por enfermedad diarreica aguda en menores de 5 años en cada una de las unidades de la federación, en el período de 1990 a 2011. Estas proporciones fueron ajustadas mediante un modelo de regresión beta, utilizando la región y el año de ocurrencia como covariables. El análisis realizado sobre el modelo corroboró los hallazgos clínicos destacados en la literatura, que demuestran la eficacia de las medidas tomadas a lo largo de los años para combatir esta enfermedad.

Palabras claves: Mortalidad Infantil, Regresión Beta, Enfermedad Diarreica Aguda, Hallazgos Clínicos, Control de la Enfermedad

1 Introduction

In Brazil, in the $mid - 1980s$, Acute Diarrheal Disease (ADD) was responsible for over 17% of infant deaths. Between 2002 and 2004, this number dropped to just over 4%, and the latest report from the Ministry of Health in 2011 showed a decrease to 1.6%. This result is due to the efforts of professionals from various sectors of society, especially those involved in public health policy practices [13, 2].

ADD is a "syndrome caused by various etiological agents whose predominant manifestation is an increase in the number of bowel movements, with watery or loose stools´´ [9]. The infection occurs via transmission of the enteropathogen, commonly through oral contact with feces [14, 15].

Worldwide, the highest rates of infant mortality from acute diarrheal diseases are in the poorest countries and regions, which is no different in Brazil [1]. In Brazil, the main factors for the spread of pathogens causing this disease are the lack of basic sanitation, malnutrition, limited and poor access to medical care and health centers, and inadequate housing [6]. Other important factors for contagion or lack of treatment often include the unavailability of free preventive vaccines for the population over six months of age [10], and the lack of maternal knowledge about the cause, symptoms, care, and especially the prevention of the disease [4].

Studies conducted in several countries reached conclusions very similar to those of Brazilian researchers. These researchers found that as social inequalities diminished or even disappeared, the mortality rates caused by ADD dropped drastically. The same was detected in Brazil, where the North and Northeast regions have the highest mortality rates from ADD, which also have the greatest number of risk factors for the spread of such diseases. The South region has the lowest mortality rate, while the Southeast and Central-West regions have intermediate rates, as confirmed by [12], which showed in their research that exclusively biological determinants were influenced by socioeconomic and demographic factors.

In this context, to develop the present research, records of the proportion of deaths caused by ADD in children under 54 years old, made available by the federal government through the DataSUS platform, were collected to evaluate whether this information statistically corroborates the empirical evidence outlined in the literature. For this purpose, a descriptive analysis of the data was performed, and a beta regression model was proposed and analyzed to verify the existence and quantify the possible relationship between the proportion of such deaths, a measure necessarily between zero and one, and the region and/or year of occurrence.

A probability distribution that encompasses the characteristics of interest is the beta distribution, used to understand the variability of a random variable Y supported on an open interval $I = (a, b) \subset \mathbb{R}$, with $a, b \in \mathbb{R}$ and $a < b$. If the random variable represents a proportion, we have the particularization where $a = 0$ and $b = 1$, that is, the unit interval $U = (0, 1) \subset \mathbb{R}$. With broad application, a regression model based on the beta distribution has already been analyzed and discussed in many studies, as seen in [5] and [8].

The beta regression model proposed by [5] has interesting characteristics in that it is analogous to those presented in the study of generalized linear models, a class already described in the literature by [7], and widely extended by the author himself and numerous others. Thus, if the aim is to model a set of observations whose response variable behavior is governed by the beta distribution and is closely related to another set of independent variables through a regression structure, the beta regression model may be the appropriate choice for this fit.

The modeling results were detailed to show that, indeed, the public policies implemented over the years were sufficiently effective in combating this type of disease and demonstrated remarkable progress in reducing the initially high proportion of deaths. The subsequent sections describe the techniques used to address the problem, the results obtained, and the analyses performed, and finally, a brief consideration of the agreement between the conclusions obtained through the statistical model and the clinical findings already described in the literature.

2 The Method

The probability density function of a random variable Y restricted to the unit interval $U = (0, 1) \in \mathbb{R}$ and governed by the beta distribution with pa-

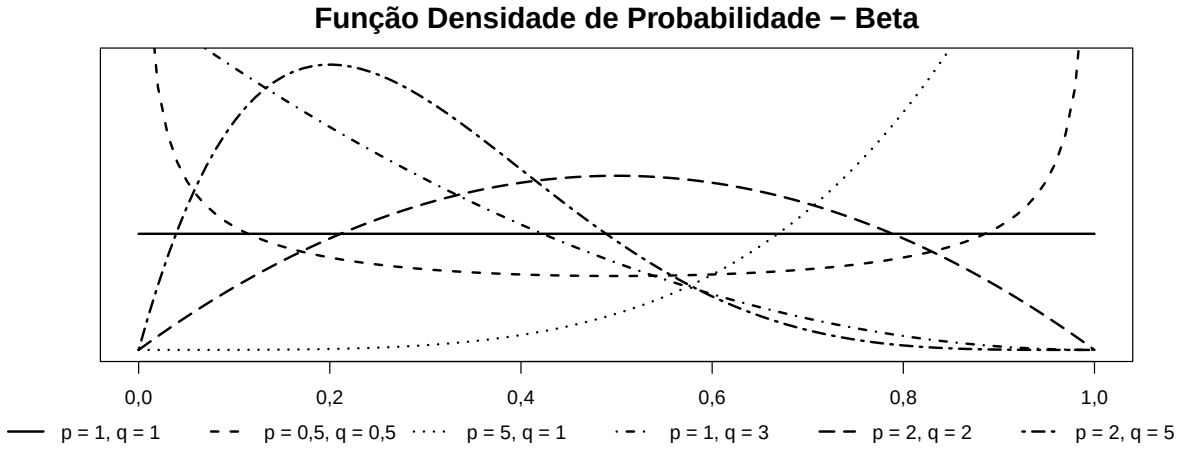

Figure 1: Some special cases for the probability density function of the beta distribution.
Source: the authors

rameters p and q, where $p, q \in \mathbb{R}_*^+$, is a real function denoted by $\mathcal{B}(p, q)$ and expressed by

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\,\Gamma(q)}\, y^{p-1}(1-y)^{q-1}, \quad y \in (0,1)\ ,$$

where the parameters p and q are positive, and $\Gamma(\cdot)$ denotes the Gamma function.

The beta distribution is one of the most versatile distributions, capable of assuming a myriad of shapes. In Figure 1, some of the *main types of shapes* can be visualized, such as the constant shape (the uniform distribution), the symmetric, unimodal, or bathtub shape (occurring when p and q are equal), the negatively skewed, unimodal, or strictly increasing shape (when $p > q$), and the positively skewed, unimodal, or strictly decreasing shape (when $p < q$).

The expected value, $\mathbb{E}(y)$, and the measure of dispersion, $\mathrm{Var}(y)$, are, respectively,

$$\mathbb{E}(y) = \frac{p}{p+q}$$

e

$$\mathrm{Var}(y) = \frac{p\,q}{(p+q)^2(p+q+1)}\ .$$

In the special cases where both parameters p and q are greater than one, the mode of the distribution exists and is given by

$$\text{Moda} = \frac{p-1}{p+q-2}\ .$$

In regression analyses, it is common and reasonably more convenient to model the expected value of the response variable, that is, it is more interesting to fit the model to the *mean* parameter. In the case of the parametrization presented earlier, the expected value of the random variable Y is a function of the parameters p and q. Furthermore, it is also reasonable to introduce a precision or dispersion parameter into the model.

The idea is to work with a new parametrization of the Beta distribution's probability density function, such that the parameters expressing the expected value and the precision are immediately identifiable. To achieve this, consider the parameters expressed by

$$\mu = \frac{p}{p+q} \qquad \text{e} \qquad \phi = p + q$$

and therefore

$$p = \mu\,\phi \qquad \text{e} \qquad q = (1-\mu)\,\phi\ ,$$

therefore, the expected value and the variance of the response variable are expressed, respectively, as

$$\mathbb{E}(y) = \mu \qquad \text{e} \qquad \mathrm{Var}(y) = \frac{\mu\,(1-\mu)}{1+\phi}\ .$$

It follows that μ is the mean parameter of the response variable and ϕ is a parameter that can be interpreted as a precision parameter, in the sense that, for a fixed value of μ, $\phi \to \infty$ implies $\mathrm{Var}(y) \to 0$, thus serving as a precision parameter. Note that the variance of the random variable Y is a function of μ, and consequently, responses

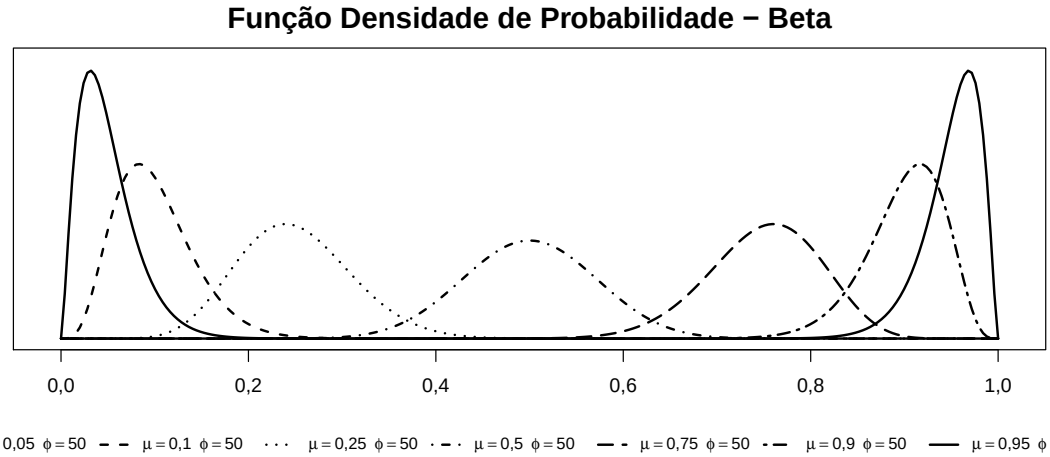

Figure 2: Some specific cases of the probability density function for the beta distribution, after reparameterization.

Source: the authors

with non-constant variances are naturally accommodated by this model.

Applying the substitutions to the probability density function of the Beta distribution, one obtains the following function, for $y \in (0,1)$

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

where the parameter ϕ is positive, $\mu \in (0,1)$, and $\Gamma(\cdot)$ denotes the Gamma function.

Observe in Figure 2 the variation in the behavior of the beta distribution's probability density function with changes in the values of the parameters μ and ϕ. Note now the clearer perception of the parameters μ and ϕ.

Note that, although the model as presented so far assumes that the random variable Y belongs to the unit interval $U = (0,1) \subset \mathbb{R}$, it is possible to extend this interval to any open $I = (a,b) \subset \mathbb{R}$, with $a < b$ and $a, b \in \mathbb{R}$ known. An extension of this form for a random variable X restricted to the interval $I = (a,b)$ is obtained through a linear transformation on X such that

$$Y = \frac{X-a}{b-a}\ ,$$

The modeling is done on Y instead of modeling X directly.

2.1 Model Definition

Consider a set of independent and identically distributed random variables $Y_1, \ldots, Y_n$, where each Y_i, with $i = 1, \ldots, n$, follows a beta distribution with mean μ_i and unknown precision ϕ. The regression model is defined by assuming that the expected value of the random variable Y_i can be written through the regression structure expressed by

$$g(\mu_i) = \sum_{j=1}^{p} x_{ij}\,\beta_j = \boldsymbol{x}_i'\boldsymbol{\beta} = \eta_i\ , \qquad p < n\ ,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ where is the vector of unknown regression parameters and$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})' \in \mathbb{R}^p$ is the vector of covariates with the set of observations of the p covariates, which take fixed and known values.

The link function g is an invertible real function, strictly monotonic, and of class C^2 that maps the unit interval $U = (0,1)$. It is the component responsible for establishing a link between the values taken by the linear predictor, η_i, and the values taken by the random variable Y_i. Some commonly used link functions among researchers, and their respective inverse functions, are as follows:

- **Link function** $\texttt{Log}(\mu_i)$**:**

$$\begin{aligned} g(\mu_i) &= \log(\mu_i) = \eta_i \\ \iff \mu_i &= e^{\eta_i} = g^{-1}(\eta_i)\ . \end{aligned}$$

- **Link function** `Logit`(μ_i)**:**

$$\begin{aligned} g(\mu_i) &= \log\left(\tfrac{\mu_i}{1-\mu_i}\right) = \eta_i \\ \iff \mu_i &= \frac{e^{\eta_i}}{e^{\eta_i}+1} = g^{-1}(\eta_i) \ . \end{aligned}$$

- **Link function** `Probit`(μ_i)**:**

$$\begin{aligned} g(\mu_i) &= \Phi^{-1}(\mu_i) = \eta_i \\ \iff \mu_i &= \Phi(\eta_i) = g^{-1}(\eta_i) \ , \end{aligned}$$

  where Φ is the cumulative distribution function of a random variable with a standard normal distribution.

- **Link function** `LogComplementar`(μ_i)**:**

$$\begin{aligned} g(\mu_i) &= \log\big(-\log(1-\mu_i)\big) = \eta_i \\ \iff \mu_i &= 1-e^{-e^{\eta_i}} = g^{-1}(\eta_i) \ . \end{aligned}$$

- **Link function** `LogLog`(μ_i)**:**

$$\begin{aligned} g(\mu_i) &= -\log\big(-\log(\mu_i)\big) = \eta_i \\ \iff \mu_i &= e^{-e^{-\eta_i}} = g^{-1}(\eta_i) \ . \end{aligned}$$

- **Link function** `Cauchy`(μ_i)**:**

$$\begin{aligned} g(\mu_i) &= \tan\left[\pi\left(\mu_i-\frac{1}{2}\right)\right] = \eta_i \\ \iff \mu_i &= \frac{1}{2}+\frac{\arctan(\eta_i)}{\pi} = g^{-1}(\eta_i) \ . \end{aligned}$$

Note that, regardless of the chosen link function, $\eta_i \to \infty$ implies $\mu_i \to 1$ and $\eta_i \to -\infty$ implies $\mu_i \to 0$. That is, large magnitudes of the linear predictor indicate that the expected value is closer to the extremes of the interval $U = (0, 1)$.

Given the above, the model used in this study is formally defined by

$$\begin{aligned} Y_i &\sim \mathcal{B}(\mu_i, \phi) \\ g_1(\mu_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} \\ g_2(\phi) &= \gamma \ , \end{aligned}$$

where the response variable Y_i and its expected value μ_i belong to the unit real interval $U = (0, 1) \subset \mathbb{R}$, the precision parameter ϕ is a positive real number, i.e., $\phi \in \mathbb{R}_*^+$, the parameter γ belongs to $\mathbb{R}$, and the parameter vectors $\boldsymbol{\beta}$ and covariates $\boldsymbol{x}$ belong to $\mathbb{R}^p$.

2.2 Estimation of Parameters

Considering an independent random sample of n observations, one can use an appropriate method to estimate the parameters $\boldsymbol{\beta}$ and ϕ. The commonly used technique consists of maximizing the likelihood function or, equivalently, the log-likelihood function, which in this case is expressed as

$$\ell(\boldsymbol{\beta}, \phi; \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \ell_i(\mu_i, \phi) \ ,$$

where

$$\begin{aligned} \ell_i(\mu_i, \phi; \boldsymbol{y}, \boldsymbol{X}) = \log\big[\Gamma(\phi)\big] - \log\big[\Gamma(\mu_i\phi)\big] - \\ \log\big[\Gamma\big((1-\mu_i)\phi\big)\big] + \\ (\mu_i\phi - 1)\log(y_i) + \\ \big[(1-\mu_i)\phi - 1\big]\log(1-y_i) \ , \end{aligned}$$

keeping in mind that

$$g(\mu_i) = \sum_{j=1}^{p} \beta_j \, x_{ij} = \eta_i \ .$$

If the number n of observations is small, it is known that the estimators of the parameters indexing the beta distribution may be highly biased. In general, the bias of the maximum likelihood estimators is of order $O(n^{-1})$ [3]. On the other hand, they have desirable properties, such as asymptotic efficiency, when considered under the assumption of large samples. If a sufficiently large sample is used, it is possible to make inferences about the regression parameters based on asymptotic theory; in this case, the commonly used statistics are the Wald, Score, and Likelihood Ratio tests [5].

2.3 Model Diagnostics

Residuals provide valuable information to assess the quality of a fit and to check if the model assumptions have been met. Numerous graphical tools are used to identify discrepancies between the fitted model and the collected observations. A balance between these two sources of information is necessary for the model to be analyzed with practical value.

To assess the goodness of fit in a beta regression model, one can use the usual diagnostic methods,

such as analyzing the Residuals *versus* Fitted Values and Residuals *versus* Observation Index plots. The behavior of the standardized residuals can be evaluated using an appropriate statistical test, a Q-Q Normal Plot, the Worm Plot as proposed by [11], or a simulated envelope plot, as described by [5]. A plot of Cook's distances can be used to detect any potential influential points, and the Residuals *versus* Linear Predictor plot can be used to assess the adequacy of the chosen link function.

3 The Data

The data on the proportions of deaths caused by DDA in children under 5 years of age were collected from the Ministry of Health's database, which is made available by the federal government through the SUS Informatics Department and can be consulted in [2]. These data correspond to all death proportions recorded between 1990 and 2011 in each of the federative units, resulting in 594 observations.

A beta regression model was fitted in an attempt to understand the behavior of the random variable Y, which denotes the proportion of deaths by DDA in children under 5 years of age, considering the presence of six covariates, one for each region of the country where the death occurred, denoted by CO (Central-West), ND (Northeast), NO (North), SD (Southeast), and SL (South); and one for the year of death, denoted by A (Year). These covariates may influence the expected value of Y, so the questions addressed here are as follows:

1. **Does the expected value of Y effectively change with the indications exerted by the covariates** CO**,** ND**,** NO**,** SD**, and** SL**?** That is, does the region where the death occurs have any significant, positive or negative, influence on the average proportion of deaths?

2. **Does the expected value of Y effectively change with the changes in the covariate** A**?** In other words, have the public policies implemented over the 22 years analyzed been efficient and/or sufficient to result in a significant reduction in the initial proportion of deaths?

3. **Is it possible to describe the random variable Y with the beta distribution, considering a regression structure that accounts for the covariates** CO**,** ND**,** NO**,** SD**,** SL**, and** A**?** Or, can the expected behavior of the proportion of this type of death be described by a specific probability distribution, taking into account the region and year of occurrence?

3.1 Descriptive analysis

To address the questions of interest, the modeling process begins with a brief descriptive analysis to identify the general behavior of the available information. Note in Figure 3 that the distribution of the frequencies of the proportion of deaths is visibly positively skewed, and if we observe the range of the sample, the mode is relatively distant from the mean, given that the observed data are restricted to the real interval $I = (0{,}003\,;\,0{,}202)$. It is natural for a greater concentration of observations to be in proportions of lower magnitude.

Note that the behavior of the frequency distribution of the empirical proportions, used to construct the histogram, is similar to one or more of the possible behaviors shown in Figures 1 and 2. This provides an indication that the beta distribution may accommodate the uncertainty of this proportion.

Now, observe Figure 4 and note that there is a relatively large variation between the observed proportions of deaths, especially for some states such as Bahia, Ceará, and Rio Grande do Norte. On the other hand, some states show very low variation, such as Alagoas, Mato Grosso do Sul, and Rio Grande do Sul. The state of occurrence of the death seems to influence both the mean and median value of this proportion as well as its variability.

Anyway, the information presented in this manner is still somewhat confusing. Perhaps it is ideal to organize these observations into larger groups, such as the five macro-regions of Brazil, and to order the proportions on a rising scale. See in Figure 5 that the evidence that the location of occurrence influences the proportion of deaths becomes evident both in the average and median values as well as in the variability. Notably, the most dispersed proportions are in the Northeast region, while the least dispersed are in the South region.

As seen in Figure 6, the passage of years seems to significantly influence the observed proportion of

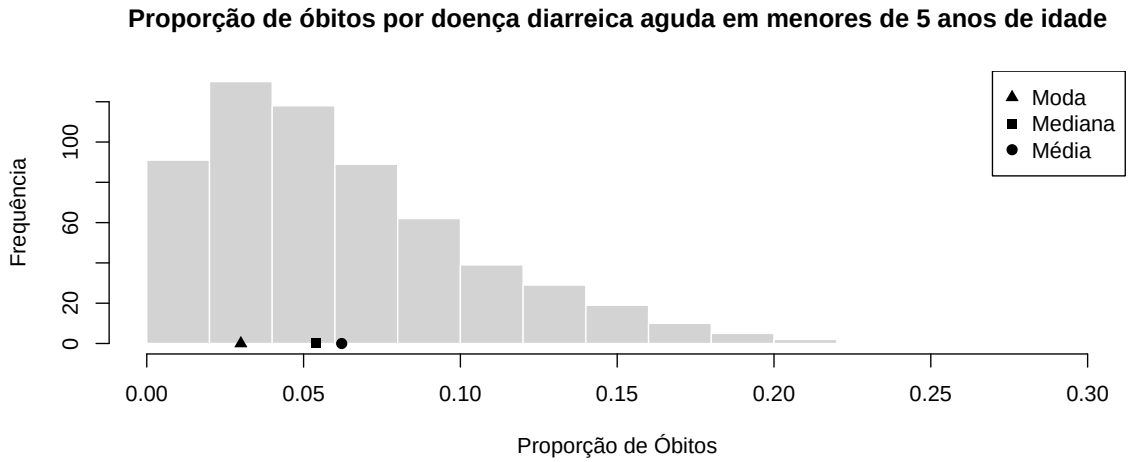

Figure 3: Histogram for the response variable Y, the proportion of deaths due to DDA in children under 5 years old.

Source: the authors

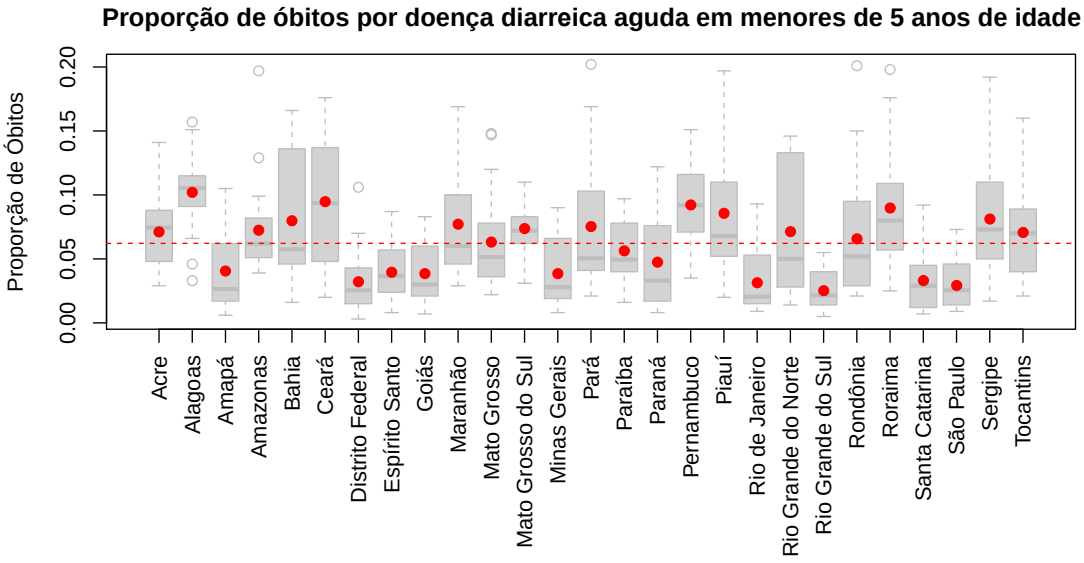

Figure 4: Box Plot for the proportion of deaths from DDA in children under 5 years of age, constructed based on the federal unit of occurrence.

Source: the authors

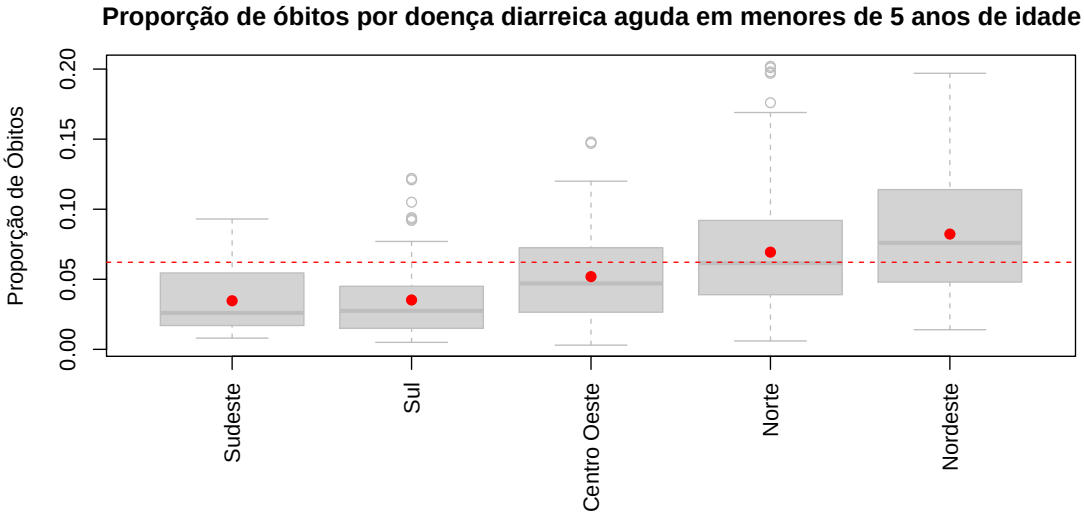

Figure 5: Box Plot for the proportion of deaths due to DDA in children under 5 years of age, constructed based on the Brazilian macro-region of occurrence.

Source: the authors

deaths. There is a reasonably progressive decrease that is evident both in the mean and median values of the proportion as well as in its variability.

The descriptive plots revealed evidence that both covariates, year and region of occurrence, affect the behavior of the death proportion

3.2 Adopted Beta Regression Model

A model formally described for the study of this problem, using the `Cauchy` link function for μ_i and the `Log` link for the parameter ϕ, is as follows:

$$
\begin{aligned}
Y_i &\sim \mathcal{B}(\mu_i, \phi) \\
\tan\left[\pi\left(\mu_i - \frac{1}{2}\right)\right] &= \beta_1 (\mathrm{CO})_i + \beta_2 (\mathrm{ND})_i + \\
&\quad \beta_3 (\mathrm{NO})_i + \beta_4 (\mathrm{SD})_i + \\
&\quad \beta_5 (\mathrm{SL})_i + \beta_6 (\mathrm{A})_i \\
\log(\phi) &= \gamma \ ,
\end{aligned}
$$

where the response variable Y_i and its expected value μ_i belong to the unit real interval $U = (0, 1) \subset \mathbb{R}$, the precision parameter ϕ is a positive real number, that is, $\phi \in \mathbb{R}_*^+$, the parameter γ belongs to $\mathbb{R}$, and the parameter vector $\boldsymbol{\beta}$ belongs to $\mathbb{R}^6$. Additionally, the covariates corresponding to the five macroregions, CO (Central-West), ND (Northeast), NO (North), SD (Southeast), and SL (South) are dichotomous variables that take the value 1 if the death observation is from the respective region, and 0 otherwise. Finally, the covariate A (Year) takes all integer values from 1991 to 2011, according to the year of the death occurrence.

3.2.1 Its adjustment

The point estimates, standard errors, and Wald statistics with the corresponding p-values for each parameter are presented in Table 1. Note that all parameters were considered statistically significant, meaning there is evidence that all regions, as well as the year of occurrence, have some influence on the proportion of deaths due to DDA in children under 5 years old.

Table 1: Estimates of the model parameters for the expected proportions of deaths due to DDA in children under 5 years old.

	Estimate	**Standard Error**	**Value t**	**Pr(> |t|)**
β_1	622,92	26,76	23,28	<2e-16
β_2	624,58	26,73	23,36	<2e-16
β_3	624,09	26,74	23,34	<2e-16
β_4	620,93	26,78	23,19	<2e-16
β_5	620,80	26,78	23,18	<2e-16
β_6	−0,31	0,01	−23,46	<2e-16
γ	−2,26	0,03	−81,56	<2e-16

Source: the authors

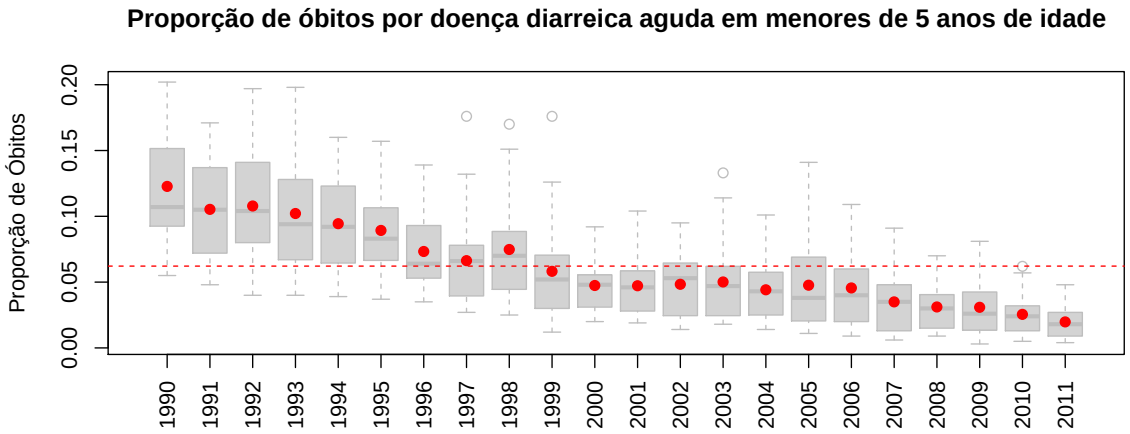

Figure 6: Box Plot for the proportion of deaths from DDA in children under 5 years of age, constructed based on the year of occurrence.

Source: the authors

The assessment of the model's adequacy, if positive, allows conclusions to be drawn based on the fit. This evaluation should be carried out through the analysis of diagnostic plots that were created using the residuals resulting from the optimization process.

3.2.2 Analysis of Residuals and Model Diagnosis

On the following two pages, see Figure 7 and note that:

- The *Residuals vs. Fitted Values* and *Residuals vs. Observation Index* plots indicate a good fit in that few observations have residuals beyond the 2 and -2 limits, respectively, which is expected. Additionally, both exhibit random behavior. In the first plot, there might be a false impression of some trend, but keep in mind that observed proportions naturally accumulate in proportions of lower magnitude, as the distribution is skewed. The concentration of points is not the main factor to observe in this plot; rather, it is the random distribution of points around zero.
- The *Estimated Density* and *Normal Q-Q Plot* graphs indicate a good fit since the estimated density and the Q-Q Plot suggest the normality of the residual quantiles computed based on the fitted model residuals.
- The *Cook's Distance* plot indicates that there are no observations with quantitatively significant influence. Note that, although some observations are visually prominent, the largest Cook's distance is slightly above 0.05, which is much lower than the thresholds commonly used for this measure. The *Residuals vs. Linear Predictor* plot, by not showing any deviations from linearity, indicates that the chosen link function does not present inadequacies that raise concerns.
- Both the *Wormplot for Residual Quantiles* and the *Half-Normal Plot of Residuals* indicate a good fit, as they show that the behavior of the residual quantiles and residuals, respectively, conforms to the expected behavior under the assumption of the validity of the fitted model.

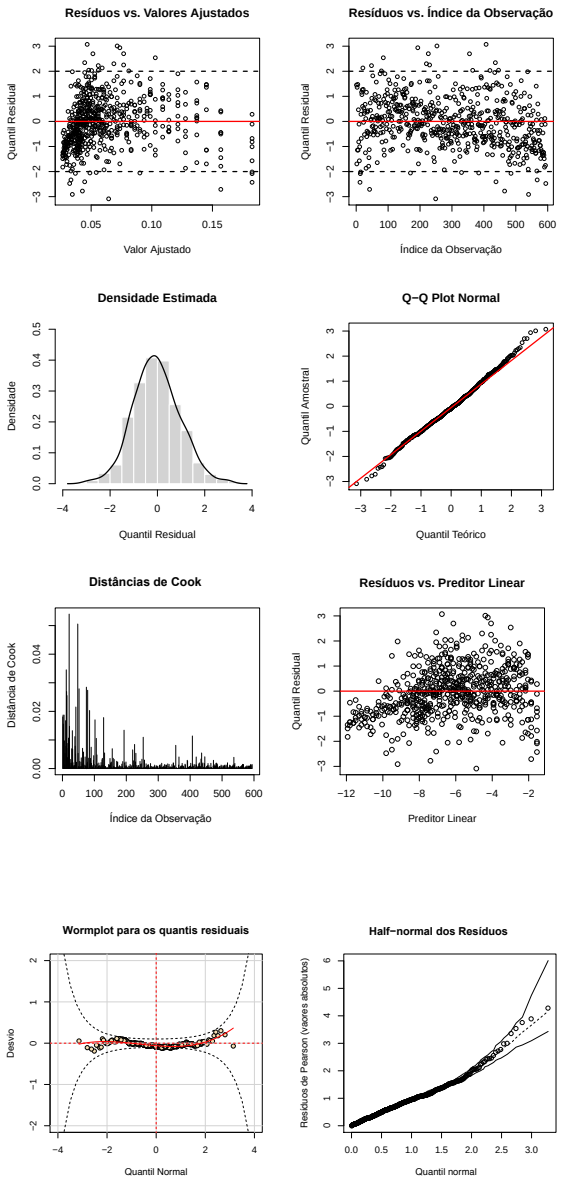

Figure 7: Gráficos diagnósticos para o modelo proposto.

Source: the authors

Based on the graphical analysis, it is assumed that the model is capable of summarizing the observed reality over the years. A pseudo-R squared of 0.4412 indicates that the model explains approximately 44.12% of the variability in the studied proportion. This is a significant number, considering the absence of additional information, as the model includes only two very broad covariates. Although this suggests the model's imprecision in providing valid estimates of the studied proportion, strictly in a quantitative sense, it is still able to capture the significance of the effects exerted by the covariates on the response variable.

A visual display of the discrepancies between the observed and model-adjusted proportions can be seen in Figure 8. The filled symbols represent the values obtained from the adjustment, while the contours represent the observed values.

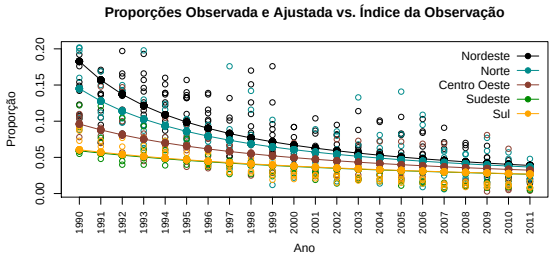

Figure 8: Observed and adjusted proportions for examining the discrepancies presented by the model.

Source: the authors

Now, with all the previous considerations, it is possible to interpret the results obtained for the estimates and understand what they represent in the context of this study. Observe Table **??** and note that:

- The estimate for the parameter β_6, related to the covariate A (Year), is negative ($\widehat{\beta}_6 = -0.31461$). This indicates that a quantitative increase in the variable A, i.e., successive progression over time, represents a decrease in the expected proportion of deaths from DDA in children under 5 years old.
- The estimates for the parameters $\beta_1, \ldots, \beta_5$, representing the covariates CO, ND, NO, SD, and SL, respectively, are all positive, which indicates that, on average, the proportion of deaths from DDA in children under 5 years old is increased according to the region where the death occurred. Additionally, the magnitude of these estimates indicates in which region(s) this average proportion is representatively higher.

Note that the values presented in Table **??** do not directly expose the information studied, as a link function was used to relate the response variable with the linear combination of the covariates. That is, if it is of interest to use the obtained es-

timates to numerically understand what they represent in the context and scale of the study, it is necessary to apply the inverse of the link function to the idealized linear combination, η_i. In this study, the following equivalence was used:

$$g(\mu_i) = \tan\left[\pi\left(\mu_i - \frac{1}{2}\right)\right] = \eta_i$$

$$\iff \mu_i = \frac{1}{2} + \frac{\arctan(\eta_i)}{\pi} = g^{-1}(\eta_i) \ .$$

In Table 2, the results obtained for the expected proportion in each region in the years 1990, 2000, and 2010 are displayed. It is possible to numerically observe the previously interpreted relationships as well as a quantified evolution of the proportion of deaths each decade. Note the significant decrease in all expected proportions over the years, regardless of the regions. The substantially higher proportion at the beginning of the 1990s for the North and Northeast regions and the subsequent approach to proportions almost equivalent to the other regions in the following decades is notable.

Table 2: Estimates for the average proportions of deaths due to DDA in children under 5 years of age over the decades.

Region	**1990**	**2000**	**2010**
Sul	0,0597	0,0376	0,0275
Sudeste	0,0611	0,0382	0,0278
Centro Oeste	0,0977	0,0501	0,0336
Norte	0,1485	0,0613	0,0383
Nordeste	0,1880	0,0676	0,0407

Source: the authors

4 Discussions

The beta regression model adequately fitted the collected data, successfully addressing the expressions of the effects considered in the study that influence the proportion of deaths caused by ADD in children under 5 years old. It was confirmed that there was a beneficial effect observed over the years, i.e., the decrease in the proportion of deaths over time. The relevance of the region where the deaths occurred on the expected proportion was also confirmed.

As a conclusion of this research, it is also necessary to highlight the success, especially in the North and Northeast regions, of the public policies implemented over the years, which aimed to reduce the proportion of deaths caused by this type of disease. These policies either established means to provide and promote family access to information and, consequently, the prevention and immediate treatment of the disease, or provided a minimal basic sanitation structure that triggered a change in the progression dynamics of these diseases.

This methodology can be extended by considering other effects that may influence the expected behavior of the proportion of deaths caused by ADD in children under 5 years old.

References

[1] BRASIL. Ministério da Saúde, Brasília. *Atenção Integrada às Doenças Prevalentes na Infância*, 1 edition, 1999.

[2] BRASIL. Ministério da Saúde. *Indicadores de mortalidade: C.6 Mortalidade proporcional por doença diarreica aguda em menores de 5 anos de idade*, 2015.

[3] Francisco Cribari-Neto and Klaus L. P. Vasconcellos. Nearly unbiased maximum likelihood estimation for the beta distribution. *J. Statist. Comput. Simul.*, 72(2):107–118, 2002.

[4] Lygia Carmen de Moraes Vanderlei and Gisélia Alves Pontes da Silva. Diarréia aguda: O conhecimento materno sobre a doença reduz o número de hospitalizações nos menores de dois anos? *Revista da Associacao Medica Brasileira*, 50(3):276–81, 2004.

[5] S. L. P. Ferrari and F. Cribari-Neto. Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, 31:799–815, 2004.

[6] A. R. FUCHS, C. G. VICTORA, and J. FACHEL. Modelo hierarquizado: Uma proposta de modelagem aplicada à investigação de fatores de risco para diarréia grave. *Revista de Saúde Pública*, 30:168–178, 1996.

[7] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):370–384, 1972.

[8] A. V. Rocha and A. B. Simas. Influence diagnostics in a general class of beta regression models. *Test*, 20:95–119, 2011.

[9] SBI. Sociedade Brasileira de Infectologia. *Doenças Diarreicas Agudas*, 2015.

[10] SUS Portal da Saúde. *Vacinação: Doença Diarreica Aguda*, mar 2014.

[11] Stef van Buuren and Miranda Fredriks. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20:1259–1277, 2001.

[12] L. C. M. Vanderlei, G. A. P. Alves, and J. U. Braga. Fatores de risco para internamento por diarréia aguda em menores de dois anos: estudo de casocontrole. *Cad Saúde Pública*, 19:455–63, 2003.

[13] C. G. Victora. Intervenções para reduzir a mortalidade infantil, pré-escolar e materna no brasil. *Revista Brasileira de Epidemiologia*, 4:3–69, 2001.

[14] Cesar G. Victora. Mortalidade por diarreia: o que o mundo pode aprender com o brasil? *Jornal de Pediatria*, 85(1):3–5, 2009.

[15] I. Zoysa, D. Carson, and R. Feachem. Perception of childhood diarrhoea and its treatment in rural zimbabwe. *Soc Sci Med*, 19:727–734, 1984.