

Statistica Sinica Preprint No: SS-2025-0172	
Title	A Class of Bootstrap Based Residuals for Compositional Data
Manuscript ID	SS-2025-0172
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0172
Complete List of Authors	Gustavo H. A. Pereira and Jianwen Cai
Corresponding Authors	Gustavo H. A. Pereira
E-mails	gpereira@ufscar.br
Notice: Accepted author version.	

A CLASS OF BOOTSTRAP BASED RESIDUALS FOR COMPOSITIONAL DATA

Gustavo H. A. Pereira^{1,2} and Jianwen Cai²

¹ *Federal University of São Carlos*, ² *University of North Carolina at Chapel Hill*

Abstract: Regression models for compositional data are common in several areas of knowledge. As in other classes of regression models, it is desirable to perform diagnostic analysis in these models using residuals that are approximately standard normally distributed. However, for regression models for compositional data, there has not been any multivariate residual that meets this requirement. In this work, we introduce a class of asymptotically standard normally distributed residuals for compositional data based on bootstrap. Monte Carlo simulation studies indicate that the distributions of the residuals of this class are well approximated by the standard normal distribution in small samples. Two other simulation studies suggest that residuals of this class can detect a wrong linear predictor, while one of them performs well in identifying an incorrect response distribution. The usefulness of this residual is illustrated through an application on sleep stages. The class of residuals proposed here can also be used in other classes of multivariate regression models.

Key words and phrases: Bootstrap, compositional data, diagnostic analysis, quantile residual.

*Authors' ORCIDs: Gustavo H. A. Pereira, 0000-0002-5491-0937
Jianwen Cai, 0000-0002-4945-6733

1. Introduction

Compositional data are proportions of disjoint categories adding to one (Hijazi and Jernigan, 2009). They are common in several areas such as medicine (Gueorguieva et al., 2008), marketing (Morais et al., 2018), ecology (Tsilimigras and Fodor, 2016), genetics (Fernandes et al., 2014), and geosciences (Filzmoser et al., 2012). This kind of data can be modeled in different ways. One possible approach is to use Dirichlet regression (Campbell and Mosimann, 1987; Hijazi and Jernigan, 2009) that can be viewed as a generalization of beta regression (Ferrari and Cribari-Neto, 2004). An overview of advances in the analysis of compositional data can be found in Alenazi (2023).

When a regression model is fitted, it is important to check model adequacy and to identify outliers and influential observations. To reach these goals, it is of interest to use residuals whose distribution is well approximated by the standard normal distribution. When the distribution of a residual is not well approximated by the standard normal distribution, it is hard to interpret residual plots. In these cases, it is not unusual that they are also nonidentically distributed and well-fitted models may be discarded (see Application 3 of Pereira (2019)). Some recent contributions related to residual analysis include Andrade et al. (2023), Li et al. (2024), Yang

(2024) and Gerber and Craig (2024), which consider residuals in settings other than compositional data.

The quantile residual (Dunn and Smyth, 1996) is simple and can be used in many regression models. It is asymptotically standard normally distributed if the parameters are consistently estimated. In addition, several papers showed that its distribution in small samples is well approximated by the standard normal distribution for different regression models (Pereira, 2019; Lemonte and Moreno-Arenas, 2019; Feng et al., 2020). Other works proposed modifications in the quantile residual to make it more suitable for some kinds of response variables (Scudilio and Pereira, 2020; Pereira et al., 2020; Andrade et al., 2023).

For compositional data, at least in principle, the quantile residual can be calculated in two different ways. One can obtain the univariate or marginal quantile residual calculating it for each component of each observation (Hijazi, 2006). The problem with this approach is that it is hard to use a univariate residual for identifying outliers and for checking model adequacy in the analysis of compositional data, because each observation has several residuals. The other option is to calculate a single multivariate quantile residual for each observation using the distribution function assumed for the compositional response variable. However, this option is not viable for

compositional data, as explained in Section 2.2.

Several other residuals were proposed for compositional data. Gueorguieva et al. (2008) describes three univariate residuals and Trijoulet et al. (2023) introduced an additional one for compositional data, but they have the same drawback of the marginal quantile residual. Gueorguieva et al. (2008) also proposed two multivariate residuals. They are known as composite residuals, because they are a function of univariate residuals. However, these residuals assume only positive values and so their distributions cannot be well approximated by the standard normal distribution.

In principle, one could use other distributions to evaluate the composite residuals. However, these residuals do not follow a chi-squared or any other known distribution, because the univariate residuals for the same observation are not independent. Therefore, for more effective diagnostic analysis, it is desirable to use a residual whose distribution can be well approximated by a known distribution.

In this work, we propose a class of multivariate bootstrap based residuals for compositional data and prove that they are asymptotically standard normally distributed. Simulation studies also suggest that their distributions are well approximated by the standard normal distribution in small samples. This class of residuals uses bootstrap to estimate the distribu-

tion function of a multivariate residual and applies the quantile function of the standard normal distribution for obtaining an asymptotically standard normally distributed residual.

An additional contribution of this work is the novel application to data on sleep stages. The proportion of time spent on each sleep stage is a compositional variable. Nonetheless, the literature typically examines each sleep stage individually, treating them as four separate proportions (Ohayon et al., 2017; Maski et al., 2021). Here, we take into account the compositional nature of this response variable.

The remainder of this paper is organized as follows. Section 2 presents the regression model considered in this work and introduces a class of multivariate bootstrap based residuals for compositional data. In Section 3, Monte Carlo simulation studies are performed to study the properties of four residuals in the proposed class and to evaluate the ability of the proposed residuals to detect model misspecification. An application to real data on sleep stages is presented in Section 4. Concluding remarks are provided in Section 5.

2. A class of bootstrap based residuals

The class of residuals proposed in this section can be used in any parametric regression model for compositional response variable and independent observations. In the simulation studies and in the applications of our work, we consider the Dirichlet regression model. Some recent contributions related to this model include Ross (2020), Hanretty (2021), Melo et al. (2022), Martínez-Minaya et al. (2023) and Ascari et al. (2024). The Dirichlet regression model has been used in applications in several areas. See some recent applications of the model, for example, in Breuninger et al. (2021), Reichert et al. (2022), Brida et al. (2023), Cudjoe et al. (2024) and Apolo et al. (2025).

2.1 Dirichlet regression model

There are two parameterizations of the Dirichlet distribution and they are known as common and alternative parameterization (Morais et al., 2018). In the former we have k shape parameters, where k is the number of components of the compositional random variable. Here we consider the alternative parametrization in which the parameters are more interpretable, and so it is more suitable for fitting regression models.

Let w be a continuous compositional random variable, i.e., $w = (w_1, w_2,$

2.1 Dirichlet regression model

$\dots, w_k)^\top$, $0 < w_j < 1$, $\sum_{j=1}^k w_j = 1$. The compositional random variable w is Dirichlet distributed with parameters $(\mu_1, \mu_2, \dots, \mu_k, \phi)^\top$, $\mu_1 = 1 - \sum_{j=2}^k \mu_j$, $0 < \mu_j < 1, \forall j \in \{1, 2, \dots, k\}$, and $\phi > 0$, if its probability density function is given by (Da-Silva and Rodrigues, 2015)

$$f(w_1, \dots, w_k; \mu_1, \dots, \mu_k, \phi) = \frac{\Gamma(\phi)}{\prod_{j=1}^k \Gamma(\phi \mu_j)} \prod_{j=1}^k w_j^{\phi \mu_j - 1}. \quad (2.1)$$

In this parameterization, $E(w_j) = \mu_j$ and $\text{Var}(w_j) = \frac{\mu_j(1-\mu_j)}{1+\phi}$. Therefore the parameter μ_j is the mean of the component j of the compositional random variable and ϕ is a precision parameter.

Let $y_1, y_2, \dots, y_n, y_i = (y_{i1}, y_{i2}, \dots, y_{ik})^\top$ be Dirichlet distributed random variables as defined in (2.1). The systematic components of the Dirichlet logistic regression model are given by

$$\begin{aligned} \log\left(\frac{\mu_{ij}}{\mu_{i1}}\right) &= x_{ij1}\beta_{j1} + x_{ij2}\beta_{j2} + \dots + x_{ijp_j}\beta_{jp_j}, \quad 2 \leq j \leq k, \\ \log(\phi_i) &= d_{i1}\gamma_1 + d_{i2}\gamma_2 + \dots + d_{ip_\phi}\gamma_{p_\phi}, \end{aligned} \quad (2.2)$$

where $(x_{ij1}, x_{ij2}, \dots, x_{ijp_j}, d_{i1}, d_{i2}, \dots, d_{ip_\phi})^\top$ are constants that represent the values of the explanatory variables and $(\beta_{j1}, \beta_{j2}, \dots, \beta_{jp_j}, \gamma_1, \gamma_2, \dots, \gamma_{p_\phi})^\top$ are unknown parameters. Note that model (2.2) uses the multinomial logit link function as is used in the multinomial logistic regression (Hosmer Jr et al., 2013). As in regression models with multinomial response (Das and Mukhopadhyay, 2014; Li et al., 2022), other link functions can be used

2.2 A class of residuals for compositional data

in regression models with Dirichlet response. However, it is convenient to use the logit link function for the equations related to the mean vector of the response variable and the logarithmic link function for the dispersion parameter, because they lead to interpretable parameters.

The parameters of model (2.2) can be estimated by maximum likelihood and the model can be fitted using the package `DirichletReg` (Maier, 2014) of software R. We used this package in the simulation studies and in the application presented in Sections 3 and 4, respectively.

2.2 A class of residuals for compositional data

For regression models for univariate responses and independent observations, the quantile residual is obtained by applying the inverse standard normal distribution function to the distribution function of the response variable which yields an asymptotically standard normally distributed residual. Extending this idea to compositional data and using the multivariate distribution function of the response variable to define a multivariate quantile residual is not adequate for two reasons. First, even if the response follows a multivariate distribution simpler than the Dirichlet distribution, the resulting residual is not necessarily a good measure of the discrepancy between an observed and a fitted value of the model. For example, consider

2.2 A class of residuals for compositional data

a fitted multivariate normal regression model and for a certain observation the response vector is (c_1, c_2) , the vector of fitted values is $(0, 0)$. If the estimated standard deviation is low, $c_1 = 0$ and c_2 is quite large, **regardless of the estimated correlation between the two components, the distribution function is close to 0.5** because it essentially corresponds to the marginal distribution function of the first component of the response. As a result, a multivariate quantile residual would be close to zero, even though its absolute value should be large. **Second, for compositional data, defining a distribution function is challenging** due to the unit-sum constraint. Even for the Dirichlet distribution that is one of the simplest distributions on the simplex, the distribution function does not have a closed form and its evaluation requires nontrivial numerical algorithms (Gouda and Szántai, 2010).

To obtain an asymptotically standard normally distributed residual that is also a good measure of the discrepancy between an observed and a fitted value of the model, we introduce three key elements. First, due to the mentioned issues with the distribution function in Dirichlet regression, we define a residual based on the distribution function of a multivariate residual. Note that the distribution function of a multivariate residual evaluated at the residual of a given observation is a good measure of the discrepancy

2.2 A class of residuals for compositional data

between an observed and a fitted value of the model. As the distributions of the multivariate residuals for compositional data are unknown, we use bootstrap (Efron and Tibshirani, 1994) to estimate the mentioned distribution function. The second element is the introduction of a sign function to avoid large absolute residual values for well-fitted observations. Finally, we use randomization to obtain a residual with a continuous distribution instead of a discrete one. These elements are explained in more detail in the remainder of this section. Therefore, we propose a class of residuals that can be obtained using the following steps.

1. For each observation, calculate $l_i = h_i r_i$, where h_i is 1 or -1 (a sign function) and r_i is a multivariate residual for compositional data. At the end of this section, we suggest two options for h_i and r_i .
2. Generate B bootstrap samples of $y = (y_1, y_2, \dots, y_n)^\top$ using the fitted model and denote each of them as $y^{\{b\}} = (y_1^{\{b\}}, y_2^{\{b\}}, \dots, y_n^{\{b\}})^\top$.
3. For each bootstrap sample, fit the compositional regression model used for the original data considering the same covariates of the original model.
4. For each bootstrap sample, obtain $l_i^{\{b\}} = h_i^{\{b\}} r_i^{\{b\}}$, where $h_i^{\{b\}}$ and $r_i^{\{b\}}$ are h_i and r_i , respectively, for the b th bootstrap sample.

2.2 A class of residuals for compositional data

5. For each observation, obtain $a_i = \#\{l_i^{\{b\}} < l_i\}$.
6. For each observation, randomly generate u_i from a uniform distribution in the interval $(a_i/(B+1), (a_i+1)/(B+1))$.
7. A residual of the proposed class is given by $s_i = \Phi^{-1}(u_i)$, where Φ is the cumulative distribution function of the standard normal distribution.

Several observations related to the described procedure are necessary. The first ones refer to step 1. As we can use as r_i any multivariate residual for compositional data and we can use different sign functions h_i , s_i is a class of residuals, because for each different kind of r_i and h_i , we have a different kind of s_i .

In compositional regression models, to define a multivariate residual, usually a univariate residual is chosen and the multivariate residual is defined as the sum of the squared univariate residuals, across the k components of each observation, as in the composite residual proposed by Gueorguieva et al. (2008). Its explicit formula is presented in Section 2.3. We use this way of obtaining a multivariate residual, and also an alternative way in which we replace the quadratic function by the absolute value function. For the univariate residual, we consider the quantile residual, also known as

2.2 A class of residuals for compositional data

pseudo residual in compositional data (Hijazi, 2006), because it is suggested as the best option to perform diagnostic analysis when the support of the response variable is the unit interval (Pereira, 2019; Lemonte and Moreno-Arenas, 2019). The univariate quantile residual for the Dirichlet regression model is defined as $q_{ij} = \Phi^{-1}[F(y_{ij}; \hat{\mu}_{ij}, \hat{\phi}_i)]$, where $\Phi(\cdot)$ and $F(\cdot)$ are the distribution functions of the standard normal distribution and of the beta distribution, respectively, and $\hat{\mu}_{ij}$ and $\hat{\phi}_i$ are the maximum likelihood estimators of μ_{ij} and ϕ_i , respectively. We denote the multivariate residual r_i as $r_i^{\{a\}}$ and $r_i^{\{q\}}$, when the absolute value function and the quadratic function are used to calculate r_i , respectively.

Another observation is related to the sign function. It is important to use a sign function h_i , because the possible choices for r_i assume only positive values and the best fitted observations have the lowest values of r_i . Therefore, if the sign function was not used, s_i would have a high absolute value for the best fitted observations, because of the usage of the function $\Phi(\cdot)$ in the step 7 of the procedure to obtain the residual. This is not an adequate behavior for a residual.

We consider in this work two different definitions for the sign function h_i . For both definitions, we use an auxiliary random variable m_i and $h_i = \text{sign}\{q_{im_i}\}$. In the first, for each observation, we consider the

2.2 A class of residuals for compositional data

sign of the univariate quantile residual from the component that has the largest distance between the observed and the fitted value (largest absolute value of the univariate quantile residual). Therefore, in this case, $m_i^{\{1\}} = \arg \max_{j \in \{1, 2, \dots, k\}} |q_{ij}|$. In the other definition, for all observations, we consider the sign of the univariate quantile residual from the component that has the largest average value of the response variable. Consequently, in this definition, $m_i^{\{2\}}$ is the same for the n observations and is given by $m_i^{\{2\}} = m = \arg \max_{j \in \{1, 2, \dots, k\}} \sum_{i=1}^n y_{ij} / n$.

The last observation is related to steps 6 and 7. It is important to consider $s_i = \Phi^{-1}(u_i)$ and not obtaining s_i as a direct function of a_i for two reasons. If we use s_i as a direct function of a_i , we could not obtain s_i when a_i equals to 0 or 1. The second reason is to make s_i continuous and allow it to converge in distribution to the standard normal distribution using a finite B . Note that, as we usually use a large B , the value of s_i is not considerably affected by the randomization made in step 6. This strategy is similar to that used to obtain the randomized quantile residual for discrete data (Dunn and Smyth, 1996).

Considering that we suggested two possible definitions for r_i and for h_i , we consider in this work four different residuals in the proposed class: $s_i^{\{a1\}}$, $s_i^{\{q1\}}$, $s_i^{\{a2\}}$, and $s_i^{\{q2\}}$. Table 1 summarizes the features of these four

2.2 A class of residuals for compositional data

residuals.

Table 1: Features of the residuals of the proposed class.

Residual	Multivariate function used	Sign function
$s_i^{\{a1\}}$	absolute value	$h_i^{\{1\}} = \text{sign}\{q_{im_i^{\{1\}}}\}, m_i^{\{1\}} = \arg \max_{j \in \{1,2,\dots,k\}} q_{ij} $
$s_i^{\{q1\}}$	quadratic	$h_i^{\{1\}} = \text{sign}\{q_{im_i^{\{1\}}}\}, m_i^{\{1\}} = \arg \max_{j \in \{1,2,\dots,k\}} q_{ij} $
$s_i^{\{a2\}}$	absolute value	$h_i^{\{2\}} = \text{sign}\{q_{im_i^{\{2\}}}\}, m_i^{\{2\}} = \arg \max_{j \in \{1,2,\dots,k\}} \sum_{i=1}^n y_{ij}/n$
$s_i^{\{q2\}}$	quadratic	$h_i^{\{2\}} = \text{sign}\{q_{im_i^{\{2\}}}\}, m_i^{\{2\}} = \arg \max_{j \in \{1,2,\dots,k\}} \sum_{i=1}^n y_{ij}/n$

As mentioned before, the following theorem holds for the residuals of the proposed class. It is proved in Section S1 of the Supplementary Material.

Theorem 1. *Let $y_1, \dots, y_n$ be compositional random variables, where each y_i has vector of parameters τ_i . If τ_i is consistently estimated, under the true model,*

$$s_i \xrightarrow{D} N(0; 1).$$

Theorem 1 guarantees that the proposed residuals are approximately standard normally distributed for large sample sizes. However, it is important to study if this approximation can be made in small samples and to evaluate their capacity to detect model misspecification. We perform these tasks in Sections 3.1 and 3.2, respectively.

2.3 Other residuals for compositional data

Note that the vector of parameters τ_i mentioned in Theorem 1 is covariate-dependent. It means that it varies across observations because it is a function of the covariates and the model parameters (regression coefficients, $(\beta_{j1}, \beta_{j2}, \dots, \beta_{jp_j}, \gamma_1, \gamma_2, \dots, \gamma_{p_\phi})^\top$). As usual in regression models, and as defined in (2.2), the regression coefficients are the same for all observations.

2.3 Other residuals for compositional data

The best known multivariate residual for compositional data is the composite residual proposed by Gueorguieva et al. (2008). It was used for example by Yoo et al. (2022) and can be calculated in the package `DirichletReg` of the software R. This multivariate residual is a sum of the squared univariate Pearson residuals across the k components of each observation. In the parameterization of the Dirichlet distribution used in this work, the composite residual is given by

$$r_i^{\{q3\}} = \sum_{j=1}^k \left(\frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\phi}_i}} \right)^2. \quad (2.3)$$

We also consider a residual similar to (2.3), but using the quantile residual as the univariate residual. The composite quantile residual is given

by

$$r_i^{\{q4\}} = \sum_{j=1}^k q_{ij}^2. \quad (2.4)$$

Residuals (2.3) and (2.4) assume only positive values and so they are not approximately standard normally distributed.

3. Simulation studies

In this section, we study the distribution of the proposed residuals in small samples and assess their ability to detect model misspecification.

3.1 Distribution of the proposed residuals in small samples

Monte Carlo simulation studies were performed to study if the distribution of the proposed residuals can be well approximated by the standard normal distribution in small samples. For all of them, we considered two values for the sample size (20 and 50), $k = 3$, $B = 1000$, and 2000 Monte Carlo replicates. In each of the scenarios evaluated, we generated two covariates, $x_{i22} = x_{i32} = d_{i2}$, $x_{i23} = x_{i33} = d_{i3}$ and $x_{i21} = x_{i31} = d_{i1} = 1$ for all i . The predictor variables were kept fixed across the 2000 replicates. All simulations were performed using the software R and the package DirichletReg. Given the goal of these simulation studies, we did not include residuals (2.3) and (2.4) in these analyses.

3.1 Distribution of the proposed residuals in small samples

In scenario 1a, the covariates were generated from the standard uniform distribution and the value of the parameters was chosen in a way that the means of μ_{i1} , μ_{i2} , and μ_{i3} across the n observations are the same and the variance of the response variable is high (ϕ close to 20). In this scenario, $\beta_{21} = -0.3$, $\beta_{22} = 1.0$, $\beta_{23} = -0.5$, $\beta_{31} = -0.3$, $\beta_{32} = -0.5$, $\beta_{33} = 1.0$, $\gamma_1 = 3.0$, and $\gamma_2 = \gamma_3 = 0$.

In scenarios 2a to 5a, we changed a single feature of the first one. In scenarios 2a and 3a, we changed the value of some parameters to make the mean of the components of the response variable far and very far from each other, respectively. In scenario 4a, we considered that ϕ is a function of the covariates using $\gamma_2 = 0.5$ and $\gamma_3 = -0.5$. In the following scenario, the first and second covariates were generated from the Bernoulli distribution with parameter 0.5 and Gamma distribution with parameters 3 and 6 (mean and variance as the standard uniform distribution), respectively. Scenarios 1b to 5b are similar to scenarios 1a to 5a, respectively, with $\gamma_1 = 4.6$ (ϕ close to 100 that corresponds to a low variance of the response variable). Table 2 summarizes the considered scenarios.

Table 3 presents the simulation results for scenario 1a and $n = 20$.

It reports the sample mean, variance, skewness, and kurtosis coefficients for $s_i^{\{a1\}}$, $s_i^{\{q1\}}$, $s_i^{\{a2\}}$, and $s_i^{\{q2\}}$ for each of the n observations across the

3.1 Distribution of the proposed residuals in small samples

Table 2: Description of the scenarios.

Scenarios	Mean of (μ_1, μ_2, μ_3)			Covariates	ϕ
1a, 1b	0.333	0.333	0.333	Uniform	Constant
2a, 2b	0.290	0.151	0.559	Uniform	Constant
3a, 3b	0.308	0.049	0.643	Uniform	Constant
4a, 4b	0.333	0.333	0.333	Uniform	Variable
5a, 5b	0.333	0.333	0.333	Bernoulli/Gamma	Constant

2000 replicates. It also includes the value of the Anderson–Darling (AD) statistic (Stephens, 1986), which is used to test whether each residual is standard normally distributed. We considered the AD statistic instead of other statistics commonly used for testing normality, because the AD test is, in general, a more powerful tool for assessing normality (Yazici and Yolacan, 2007). Note that we also report the mean of the μ_i values, since the departure of a residual distribution from the standard normal distribution may depend on the values of the mean response vector. (Pereira, 2019). The mean and the variance of the residuals are very close to zero and one, respectively, for all observations and proposed residuals. The skewness coefficient is also very close to zero for almost all observations and residuals

3.1 Distribution of the proposed residuals in small samples

and not so far from zero for the remaining observations. The kurtosis coefficients of the residuals are farther from the corresponding value of the standard normal distribution than the other measures, but **none is** far from 3 for any residual. **Consequently**, the values of the AD statistics are all small and the distributions of the four proposed residuals **are** well approximated by the standard normal distribution in this scenario with $n = 20$. The quality of this approximation **is** similar for the four residuals since the mean AD statistics are very close to each other.

Table 4 presents the average values of the mean, variance, skewness, kurtosis coefficient, and AD **statistics** for all proposed residuals **across** the ten studied scenarios with $n = 20$. The **distributions** of the four proposed residuals **are** well approximated by the standard normal distribution in all scenarios when $n = 20$, **as all average** measures for the four proposed residuals are very close to the theoretical values of this distribution. The average values of the AD statistics are small for the four residuals in the ten scenarios, suggesting that this approximation is not affected by the value of the parameters related to the mean of the response variable, nor by the variance of the response nor by the distribution of the covariates. The quality of the standard normality approximation **is** similar for the four residuals in all scenarios. Results for $n = 50$ are similar and are presented

3.1 Distribution of the proposed residuals in small samples

Table 3: Simulation results Scenario 1a - $n = 20$.

i	Mean				Variance				Skewness				Kurtosis				AD Statistic			
	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$
1	0.01	0.01	0.01	0.01	0.94	0.94	0.95	0.94	0.03	0.06	0.00	-0.03	2.96	3.09	3.02	3.01	0.88	0.94	0.76	0.82
2	-0.01	-0.02	0.02	0.02	1.06	1.06	1.06	1.06	0.13	0.10	-0.01	-0.01	2.96	2.88	2.98	2.93	1.43	1.45	1.23	1.21
3	0.01	0.01	0.04	0.04	1.05	1.05	1.06	1.06	-0.07	-0.07	-0.02	0.01	3.09	3.16	3.20	3.22	0.72	0.83	1.63	1.77
4	-0.02	-0.02	-0.01	-0.01	0.97	0.98	0.97	0.98	-0.06	-0.04	0.03	0.03	2.91	2.90	2.90	2.90	0.52	0.49	0.46	0.40
5	0.02	0.02	-0.01	-0.01	0.98	0.96	0.98	0.97	0.11	0.10	-0.04	-0.03	3.09	2.99	3.13	3.05	1.19	1.16	0.74	0.75
6	0.00	0.00	-0.01	-0.01	0.99	0.99	0.99	0.99	0.05	0.05	-0.03	-0.04	3.31	3.37	3.40	3.40	1.09	1.21	1.14	1.25
7	0.01	0.01	-0.01	-0.01	1.04	1.04	1.04	1.04	0.04	0.00	-0.17	-0.14	3.13	3.17	3.10	3.09	0.67	0.75	0.98	0.92
8	0.00	0.00	-0.03	-0.04	1.02	1.02	1.02	1.02	0.03	0.03	0.01	-0.03	3.02	2.99	3.03	2.98	0.42	0.39	1.13	1.30
9	0.03	0.04	-0.01	-0.01	1.02	1.03	1.03	1.03	0.09	0.11	0.01	0.00	3.06	3.11	3.14	3.08	1.17	1.20	0.47	0.40
10	0.00	-0.01	0.02	0.02	1.02	1.02	1.02	1.02	0.08	0.06	-0.06	-0.06	2.99	3.02	3.04	3.07	0.61	0.58	0.92	0.92
11	-0.02	-0.03	0.01	0.01	0.95	0.95	0.95	0.95	0.05	0.02	-0.07	-0.06	3.00	3.03	3.06	3.08	1.37	1.38	0.86	0.89
12	-0.02	-0.02	0.02	0.02	0.99	0.99	0.99	0.99	0.00	-0.01	-0.04	-0.09	2.98	2.95	2.90	2.94	0.46	0.50	0.86	0.91
13	0.01	0.00	0.02	0.02	1.02	1.03	1.03	1.03	0.01	-0.01	0.01	-0.01	2.94	2.96	2.92	2.94	0.71	0.63	0.88	0.86
14	0.04	0.03	-0.03	-0.03	0.99	0.99	1.00	1.01	-0.03	-0.03	0.02	0.01	3.01	3.03	3.05	3.11	1.85	1.90	1.04	1.01
15	0.03	0.03	-0.04	-0.04	1.00	1.00	1.01	1.00	0.08	0.07	-0.01	-0.02	2.91	2.89	3.02	2.95	0.95	0.91	1.39	1.44
16	0.01	0.01	-0.01	-0.01	0.97	0.98	0.98	0.98	0.07	0.03	-0.03	-0.01	2.82	2.84	2.89	2.86	0.30	0.21	0.30	0.23
17	-0.02	-0.03	0.02	0.01	0.99	0.98	0.98	0.98	0.13	0.09	0.03	0.00	3.26	3.01	3.02	3.03	1.13	1.08	0.81	0.79
18	-0.02	-0.02	0.00	0.00	1.03	1.04	1.03	1.03	-0.05	-0.02	-0.04	-0.03	2.95	2.93	2.87	2.94	0.70	0.65	0.41	0.43
19	0.02	0.02	0.00	0.00	0.94	0.95	0.96	0.96	0.01	-0.03	-0.05	-0.07	2.94	3.04	3.01	3.07	1.01	0.95	0.54	0.53
20	0.00	0.00	-0.04	-0.04	1.02	1.02	1.02	1.02	0.07	0.05	-0.03	-0.02	2.90	2.89	2.97	2.94	0.67	0.62	1.76	1.95
Mean	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	0.04	0.03	-0.03	-0.03	3.01	3.01	3.03	3.03	0.89	0.89	0.92	0.94
SD	0.02	0.02	0.02	0.02	0.03	0.04	0.03	0.03	0.06	0.05	0.05	0.04	0.12	0.13	0.12	0.12	0.39	0.41	0.39	0.45

3.2 Detecting model misspecification

in Table S1 of the Supplementary Material.

We also considered a setting similar to scenario 1a, but using $k = 5$ and $k = 9$. For both values of k , and for j ranging from 2 to 5 when $k = 5$ and from 2 to 9 when $k = 9$, we considered $\beta_{j1} = -0.3, \beta_{j2} = 1.0, \beta_{j3} = -0.5$ for even values of j and $\beta_{j1} = -0.3, \beta_{j2} = -0.5, \beta_{j3} = 1.0$ for odd values of j . The results are presented in Tables S2 and S3 of the Supplementary Material. As observed for $k = 3$, the distributions of the four residuals of the proposed class for $k = 5$ and $k = 9$ are well approximated by the standard normal distribution.

3.2 Detecting model misspecification

We assess the residuals' ability to detect model misspecification related to two different aspects. In Section S3.1 of the Supplementary Material, we evaluate whether the residuals can detect misspecification in the distribution of the response. Later (Section S3.2), we study the residuals' ability to detect model misspecification in the linear predictor.

The main conclusion of the first analysis is that $s_i^{\{a1\}}$ is effective in detecting misspecification of the assumed response distribution and performs better than the other residuals considered in this paper. We also concluded that two other residuals of the proposed class can identify omitted covari-

3.2 Detecting model misspecification

Table 4: Average distributional measures for the residuals - $n = 20$.

Scenarios		$\gamma_1 = 3.0$				$\gamma_1 = 4.6$			
		$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$	$s_i^{\{a1\}}$	$s_i^{\{q1\}}$	$s_i^{\{a2\}}$	$s_i^{\{q2\}}$
1	Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Variance	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness	0.04	0.03	-0.03	-0.03	0.00	-0.01	0.01	0.00
	Kurtosis	3.01	3.01	3.03	3.03	2.98	3.01	2.96	2.99
	A-D Statistics	0.89	0.89	0.92	0.94	0.90	0.87	0.92	0.92
2	Mean	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	0.00
	Variance	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness	0.01	0.00	0.01	0.01	0.00	0.00	-0.01	-0.01
	Kurtosis	2.99	3.00	3.00	2.99	2.99	2.99	3.01	3.01
	A-D Statistics	0.91	0.90	0.98	0.99	0.99	0.99	0.80	0.80
3	Mean	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00
	Variance	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness	-0.01	-0.01	0.00	0.01	0.00	0.00	0.00	0.00
	Kurtosis	2.99	3.00	2.97	2.97	3.01	3.02	2.99	3.01
	A-D Statistics	1.24	1.27	0.81	0.79	0.84	0.83	0.86	0.81
4	Mean	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00
	Variance	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness	0.01	0.00	-0.01	0.00	0.00	0.00	0.00	0.00
	Kurtosis	2.95	2.98	2.99	3.00	2.99	3.01	2.97	2.99
	A-D Statistics	1.46	1.46	1.22	1.18	0.91	0.89	1.17	1.11
5	Mean	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	Variance	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Skewness	0.01	0.01	-0.01	-0.01	0.00	0.00	-0.01	-0.02
	Kurtosis	3.00	3.01	3.03	3.01	2.97	2.97	2.97	2.97
	A-D Statistics	0.65	0.66	0.90	0.91	1.08	1.05	0.88	0.85

ates in the model. On the other hand, existing residuals cannot detect model misspecification. Further details of these analyses can be found in Section S3 of the Supplementary Material.

4. Application to data on sleep stages

The human sleep has four stages: N1, N2, N3 and REM. During the first two stages, the person is in a light sleep (Ritkala-Castren et al., 2015). The stage N3 corresponds to deep sleep and it is when muscles and tissues are healed (Hussain et al., 2022). Finally, the stage REM is characterized by rapid eye movements and fast breathing and this stage is responsible for optimal emotional processing (Tempesta et al., 2018).

Considering the functions of the last two stages, it is important that people spend a considerable percentage of their sleep on stages N3 and REM. Sleep is considered as within normal parameters when the percentage of time spent on the stages N1, N2, N3, and REM are on the following ranges: 3 – 8%, 45 – 55%, 15 – 20%, 20 – 25%, respectively (González-Naranjo et al., 2019).

The proportion of time spent on each sleep stage is a compositional variable. However, as mentioned earlier, papers usually study each of the stages separately considering four independent proportions. To the best of

our knowledge, this paper is the first that considers the proportion of time spent on each sleep stage as a compositional variable.

Veje et al. (2021) collected sleep data from 42 adults, in which 22 were patients with Tick-borne encephalitis (TBE) and 20 controls. Their main goal was to investigate if the TBE affects the quality of the sleep measured by a questionnaire. Here, we use a Dirichlet regression model to study if the proportion of time spent on each sleep stage changes with the total sleep time (TST) and from patients with TBE to healthy adults. The sample means for the proportion of time spent on the N1, N2, N3 and REM stages are 0.134, 0.532, 0.152, and 0.182, respectively. These sample means are similar for TBE adults and controls and the standard deviations are less than 0.1.

The Dirichlet regression model assumes that $P(y_{ij} = 0) = 0$ for all i and j . As a result, the model can not be fitted if there are one or more zeros in the response variable. When there are many zeros, a possible solution is to fit a zero adjusted Dirichlet regression model (Tsagris and Stewart, 2018). However, the parameters of the discrete component of this model can not be well estimated when there are few zeros. In this case, it is better to transform data in a way that eliminates the zeros (Hijazi, 2011). When there are very few zeros, it is reasonable to simply replace the zeros by

a small value and then normalizing the imputed compositions (Aitchison, 1986).

Two patients of the data collected by Veje et al. (2021) do not spend time on the N3 stage in their monitored night of sleep. We replaced their zero values by 0.001 and normalized the proportions of these patients to sum 1.

We fitted a Dirichlet regression model for the transformed data containing the mentioned covariates in all submodels. Using the log likelihood ratio test, we concluded that there is no evidence that the parameters of the proportion of time spent on each sleep stage changes from patients with TBE to healthy adults ($p = 0.7469$). However, the mean vector ($p = 0.0019$) and the precision parameter ($p = 0.0030$) of the response variable are functions of the total sleep time.

As the residual $s_i^{\{a1\}}$ performed better in the analysis for detecting misspecification in the distribution of the response, we used it to conduct the diagnostic analysis. Figure 1 presents a normal probability plot with simulated envelope for $s_i^{\{a1\}}$ considering the final model. The plot does not suggest model misspecification of the Dirichlet regression model for the proportion of time spent on each sleep stage with the total sleep time as a covariate of all submodels.

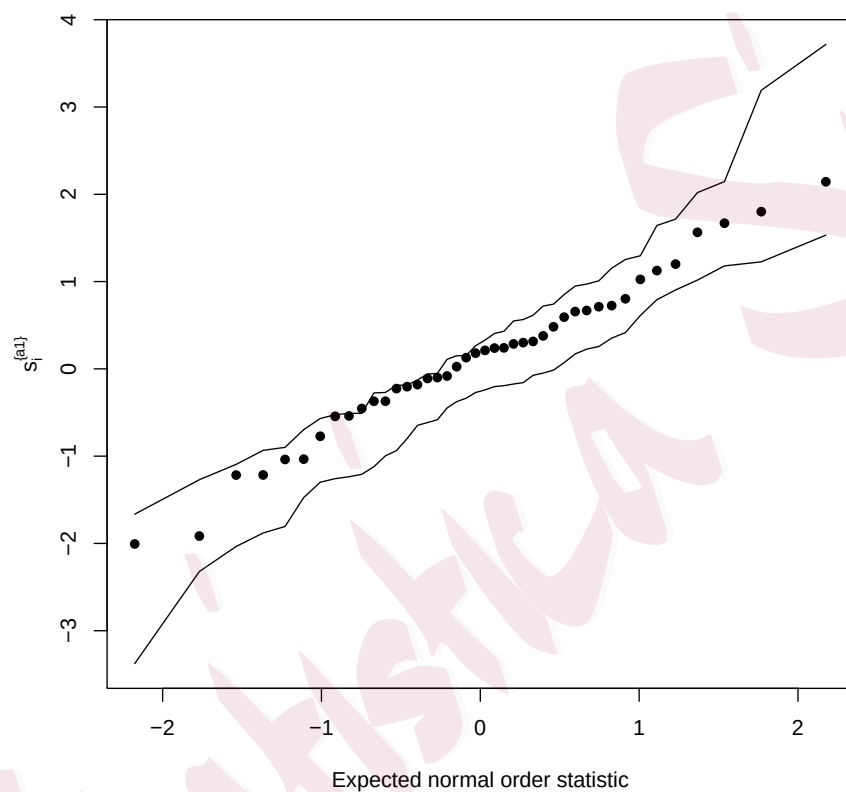


Figure 1: Normal probability plot with simulated envelope for s_i^{a1} considering the final model for data on sleep stages

Table 5 presents the parameter estimates and standard errors for the final model. Note that the estimates of the parameters associated with TST in the second and third submodels are positive. Therefore, on average, an increase in the total sleep time leads to a higher proportion of time spent on stages N3 and REM in relation to the N1 stage. These results suggest that adults have better sleep quality if they spend enough time on sleeping.

Calculating the exponential of the parameter estimates (last column of Table 5), it is estimated that for every hour increase in the total sleep time, the ratio between the mean of the proportion of time spent on the N3 stage of sleep and the mean of the proportion of time spent on the N1 stage increases by 40.3%. Similarly, it is estimated that for every hour increase in the total sleep time, the ratio between the mean of the proportion of time spent on the REM stage of sleep and the mean of the proportion of time spent on the N1 stage increases by 23.8%.

In this application on sleep stages, as mentioned before, we replaced two zero values by $\epsilon = 0.001$. To investigate if the choice of the value of ϵ changes the main conclusions presented in this section, we considered two additional values of ϵ (0.01 and 0.0001). The main conclusions are the same for the three values of ϵ . In addition, the estimates and standard errors, in general, do not change substantially for these three values of ϵ . This

Table 5: Parameter estimates and standard errors in the model for the proportion of time spent on each sleep stage.

Submodel	Covariate	Estimate	Std. Error	P-value	Exp(estim)
μ_{N2}	Intercept	0.5142	0.5949	0.3870	1.672
	TST	0.1314	0.0880	0.1350	1.140
μ_{N3}	Intercept	-2.2371	0.7087	0.0016	0.107
	TST	0.3386	0.1030	0.0010	1.403
μ_{REM}	Intercept	-1.0275	0.6688	0.1244	0.358
	TST	0.2136	0.0981	0.0295	1.238
ϕ	Intercept	0.8401	0.7433	0.2584	2.317
	TST	0.3400	0.1139	0.0028	1.405

analysis suggests that the results are not sensitive to the choice of ϵ in this application.

5. Concluding remarks

In this work, we proposed a class of asymptotically standard normally distributed residuals for compositional data based on bootstrap. We investigated the properties of four residuals of the proposed class in small samples

and the ability of the proposed residuals to detect model misspecification using Monte Carlo simulation studies. We also presented an application on sleep stages that illustrates the usefulness of one of the residuals of the proposed class.

To facilitate the diagnostic analysis of a regression model, it is desirable to find a residual whose distribution is well approximated by the standard normal distribution. To fill in the gap in the literature for compositional data, we proposed a class of residuals and proved that the residuals of the proposed class are asymptotically standard normally distributed, and hence their distribution is close to the standard normal distribution when the sample size is large. In addition, simulation studies suggested that the distributions of the four investigated residuals of the proposed class are very close to the standard normal distribution even when $n = 20$ for all considered scenarios.

It is also important that a residual can detect model misspecification. A simulation study suggested that the residual of the proposed class denoted by $s_i^{\{a1\}}$ is good for identifying that the assumed response distribution is wrong and better than the other residuals considered in this paper. In addition, another simulation study suggested that two other residuals of the proposed class can effectively identify when a covariate is missing from

the model.

Another interesting feature of the proposed class of residuals is that their expressions do not depend on the assumed distribution of the response variable. The algebraic expression of the multivariate residuals based on the deviance residual, for example, have to be obtained for each different distribution assumed for the response variable.

In addition to the Dirichlet regression model, there are several others commonly used for compositional data. Most of them are based on transformations, and the logistic-normal model (Aitchison, 1986) is the most well-known in this class. To use the proposed residual for the logistic-normal model, it is only necessary to use a r_i that is suitable for multivariate normal regression. For example, one can use the Mahalanobis distance (Ghorbani, 2019). The other steps of the algorithm described in Section 2 can be performed in the same way as in the Dirichlet regression model. For other models for compositional data, it is also only necessary to use a suitable r_i and the same algorithm to obtain a residual of the proposed class.

It is not unusual that compositional data have many zeros in one or more components of the response. In this case, it should be modeled using a regression model suitable for this situation (see for example Scealy and Welsh (2011) and Tsagris and Stewart (2018)). The proposed class

of residuals can be used, with a small modification, when data have many zeros. We have to change the univariate residual used to obtain r_i from the quantile residual to an alternative suitable for response variables that are discrete at zero and continuous in a certain interval. A possible option is the randomized quantile residual (Dunn and Smyth, 1996). Future work can also investigate when to prefer using the zero-adjusted Dirichlet regression model instead of the Dirichlet regression model in cases where the number of zeros is neither very low nor large.

Besides the usage for diagnostic analysis in regression models for compositional data, the proposed class of residuals can also be used in other classes of multivariate regression models. There is no approximately standard normally distributed multivariate residual in the literature for multivariate regression models. Therefore, our class of residuals can be a useful tool for several classes of multivariate regression models. For any multivariate regression model, a residual from the proposed class can be obtained by choosing an appropriate r_i and applying the algorithm described in Section 2. Future work can investigate the properties of residuals of the proposed class in different multivariate regression models, such as the multivariate normal regression model (Johnson and Wichern, 2013), the multivariate elliptical regression model (Lemonte and Patriota, 2011), and the multivariate

REFERENCES

generalized linear model (Schaid et al., 2019).

Supplementary Material

The Supplementary Material includes the proof of Theorem 1 (Section S1), three tables from the simulation studies presented in Section 3.1 (Section S2), and analyses of misspecification detection (Section S3).

Acknowledgements

This work was partially supported by São Paulo Research Foundation (FAPESP), grant number 2020/16334-9. We also thank an associate editor and two anonymous reviewers for their helpful comments.

References

- Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall.
- Alenazi, A. (2023). A review of compositional data analysis and recent advances. *Communications in Statistics-Theory and Methods* 52(16), 5535–5567.
- Andrade, A. C., G. H. Pereira, and R. Artes (2023). The circular quantile residual. *Computational Statistics & Data Analysis* 178, 107612.
- Apolo, A. B., C. Michaels-Igbokwe, N. I. Simon, D. J. Benjamin, M. Farrar, Z. Hepp, L. Mucha, S. Heidenreich, K. Cutts, N. Krucien, et al. (2025). Patient preferences for first-line treat-

REFERENCES

- ment of locally advanced or metastatic urothelial carcinoma: an application of multidimensional thresholding. *The Patient-Patient-Centered Outcomes Research* 18(1), 77–87.
- Ascari, R., A. M. Di Brisco, S. Migliorati, and A. Ongaro (2024). A multivariate mixture regression model for constrained responses. *Bayesian Analysis* 19(2), 377–405.
- Breuninger, T. A., N. Wawro, J. Breuninger, S. Reitmeier, T. Clavel, J. Six-Merker, G. Pestoni, S. Rohrmann, W. Rathmann, A. Peters, et al. (2021). Associations between habitual diet, metabolic disease, and the gut microbiota using latent dirichlet allocation. *Microbiome* 9(1), 1–18.
- Brida, J. G., B. Lanzilotta, and L. Moreno (2023). Compositional tourists’ expenditure: Modeling through dirichlet regression. *Tourism Economics* 29(6), 1442–1460.
- Campbell, G. and J. Mosimann (1987). Multivariate methods for proportional shape. In *ASA Proceedings of the Section on Statistical Graphics*, Volume 1, pp. 10–17. Washington DC.
- Cudjoe, E., F. Bravo, and R. Ruiz-Peinado (2024). Allometry and biomass dynamics in temperate mixed and monospecific stands: Contrasting response of scots pine (*pinus sylvestris* l.) and sessile oak (*quercus petraea* (matt.) Liebl.). *Science of the Total Environment* 953, 176061.
- Da-Silva, C. Q. and G. S. Rodrigues (2015). Bayesian dynamic dirichlet models. *Communications in Statistics-Simulation and Computation* 44(3), 787–818.
- Das, I. and S. Mukhopadhyay (2014). On generalized multinomial models and joint percentile estimation. *Journal of Statistical Planning and Inference* 145, 190–203.

REFERENCES

- Dunn, P. K. and G. K. Smyth (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* 5(3), 236–244.
- Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.
- Feng, C., L. Li, and A. Sadeghpour (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology* 20(1), 1–21.
- Fernandes, A. D., J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell, and G. B. Gloor (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing rna-seq, 16s rna gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* 2(1), 1–13.
- Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics* 31(7), 799–815.
- Filzmoser, P., K. Hron, and C. Reimann (2012). Interpretation of multivariate outliers for compositional data. *Computers & Geosciences* 39, 77–85.
- Gerber, E. A. and B. A. Craig (2024). Residuals and diagnostics for multinomial regression models. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 17(1), e11645.
- Ghorbani, H. (2019). Mahalanobis distance and its application for detecting multivariate outliers. *Facta Universitatis, Series: Mathematics and Informatics*, 583–595.

REFERENCES

- González-Naranjo, J. E., M. Alfonso-Alfonso, D. Grass-Fernandez, L. M. Morales-Chacón, I. Pedroso-Ibáñez, Y. Ricardo-De La Fe, and A. Padrón-Sánchez (2019). Analysis of sleep macrostructure in patients diagnosed with parkinson’s disease. *Behavioral Sciences* 9(1), 6.
- Gouda, A. A. and T. Szántai (2010). On numerical calculation of probabilities according to dirichlet distribution. *Annals of Operations Research* 177(1), 185–200.
- Gueorguieva, R., R. Rosenheck, and D. Zelterman (2008). Dirichlet component regression and its applications to psychiatric data. *Computational statistics & data analysis* 52(12), 5344–5355.
- Hanretty, C. (2021). Forecasting multiparty by-elections using dirichlet regression. *International Journal of Forecasting* 37(4), 1666–1676.
- Hijazi, R. (2011). An em-algorithm based method to deal with rounded zeros in compositional data under dirichlet models. In *Proceedings of CoDaWork’11: 4th international workshop on Compositional Data Analysis*, Egozcue, JJ, Tolosana-Delgado, R. and Ortego, MI (eds.) 2011. CIMNE.
- Hijazi, R. H. (2006). Residuals and diagnostics in dirichlet regression. *ASA Proceedings of the General Methodology Section*, 1190–1196.
- Hijazi, R. H. and R. W. Jernigan (2009). Modelling compositional data using dirichlet regression models. *Journal of Applied Probability & Statistics* 4(1), 77–91.
- Hosmer Jr, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied logistic regression* (3

REFERENCES

- ed.). John Wiley & Sons.
- Hussain, I., M. A. Hossain, R. Jany, M. A. Bari, M. Uddin, A. R. M. Kamal, Y. Ku, and J.-S. Kim (2022). Quantitative evaluation of eeg-biomarkers for prediction of sleep stages. *Sensors* 22(8), 3079.
- Johnson, R. A. and D. W. Wichern (2013). *Applied multivariate statistical analysis* (6 ed.). Prentice hall Upper Saddle River, NJ.
- Lemonte, A. J. and G. Moreno-Arenas (2019). On residuals in generalized johnson sb regressions. *Applied Mathematical Modelling* 67, 62–73.
- Lemonte, A. J. and A. G. Patriota (2011). Multivariate elliptical models with general parameterization. *Statistical Methodology* 8(4), 389–400.
- Li, D., M. F. Al-Mahamda, Y. Song, S. Feng, and N. N. Sze (2022). An alternate crash severity multicategory modeling approach with asymmetric property. *Analytic Methods in Accident Research* 35, 100218.
- Li, W., D. Cook, E. Tanaka, and S. VanderPlas (2024). A plot is worth a thousand tests: Assessing residual diagnostics with the lineup protocol. *Journal of Computational and Graphical Statistics* 33(4), 1497–1511.
- Maier, M. J. (2014). Dirichletreg: Dirichlet regression for compositional data in r. *Research Report Series / Department of Statistics and Mathematics* 125.
- Martínez-Minaya, J., F. Lindgren, A. López-Quílez, D. Simpson, and D. Conesa (2023). The

REFERENCES

- p>integrated nested laplace approximation for fitting dirichlet regression models.
- Journal of Computational and Graphical Statistics*
- 32(3), 805–823.
- Maski, K. P., A. Colclasure, E. Little, E. Steinhart, T. E. Scammell, W. Navidi, and C. Diniz Behn (2021). Stability of nocturnal wake and sleep stages defines central nervous system disorders of hypersomnolence. *Sleep* 44(7), zsab021.
- Melo, T. F., T. M. Vargas, A. J. Lemonte, and G. Moreno-Arenas (2022). Higher-order asymptotic refinements in the multivariate dirichlet regression model. *Communications in Statistics-Simulation and Computation* 51(1), 53–71.
- Morais, J., C. Thomas-Agnan, and M. Simioni (2018). Using compositional and dirichlet models for market share regression. *Journal of Applied Statistics* 45(9), 1670–1689.
- Ohayon, M., E. M. Wickwire, M. Hirshkowitz, S. M. Albert, A. Avidan, F. J. Daly, Y. Dauvilliers, R. Ferri, C. Fung, D. Gozal, et al. (2017). National sleep foundation’s sleep quality recommendations: first report. *Sleep health* 3(1), 6–19.
- Pereira, G. H., J. Scudilio, M. Santos-Neto, D. A. Botter, and M. C. Sandoval (2020). A class of residuals for outlier identification in zero adjusted regression models. *Journal of Applied Statistics* 47(10), 1833–1847.
- Pereira, G. H. A. (2019). On quantile residuals in beta regression. *Communications in Statistics-Simulation and Computation* 48(1), 302–316.
- Reichert, S., V. Berger, D. J. F. Dos Santos, M. Lahdenperä, U. K. Nyein, W. Htut, and V. Lummaa (2022). Age related variation of health markers in asian elephants. *Experi-*

REFERENCES

- mental Gerontology* 157, 111629.
- Ritmala-Castren, M., I. Virtanen, S. Leivo, K.-M. Kaukonen, and H. Leino-Kilpi (2015). Sleep and nursing care activities in an intensive care unit. *Nursing & health sciences* 17(3), 354–361.
- Ross, G. J. (2020). Tracking the evolution of literary style via dirichlet–multinomial change point regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183(1), 149–167.
- Scealy, J. L. and A. Welsh (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 73(3), 351–375.
- Schaid, D. J., X. Tong, A. Batzler, J. P. Sinnwell, J. Qing, and J. M. Biernacka (2019). Multi-variate generalized linear model for genetic pleiotropy. *Biostatistics* 20(1), 111–128.
- Scudilio, J. and G. H. Pereira (2020). Adjusted quantile residual for generalized linear models. *Computational Statistics* 35(1), 399–421.
- Stephens, M. A. (1986). Tests based on edf statistic. In *Goodness-of-fit Techniques*, pp. 97–193. Marcel Dekker.
- Tempesta, D., V. Socci, L. De Gennaro, and M. Ferrara (2018). Sleep and emotional processing. *Sleep medicine reviews* 40, 183–195.
- Trijoulet, V., C. M. Albertsen, K. Kristensen, C. M. Legault, T. J. Miller, and A. Nielsen (2023).

REFERENCES

- Model validation for compositional data in stock assessment models: calculating residuals with correct properties. *Fisheries Research* 257, 106487.
- Tsagris, M. and C. Stewart (2018). A dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics* 39(3), 398–412.
- Tsilimigras, M. C. and A. A. Fodor (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of epidemiology* 26(5), 330–335.
- Veje, M., M. Studahl, E. Thunström, E. Stentoft, P. Nolskog, Y. Celik, and Y. Peker (2021). Sleep architecture, obstructive sleep apnea and functional outcomes in adults with a history of tick-borne encephalitis. *PLoS One* 16(2), e0246767.
- Yang, L. (2024). Double probability integral transform residuals for regression models with discrete outcomes. *Journal of Computational and Graphical Statistics* 33(3), 787–803.
- Yazici, B. and S. Yolacan (2007). A comparison of various tests of normality. *Journal of statistical computation and simulation* 77(2), 175–183.
- Yoo, J., M. Greenacre, and D. Chung (2022). A guideline for the statistical analysis of compositional data in immunology. *Communications for Statistical Applications and Methods* 29(4), 453–469.
- Gustavo H. A. Pereira, Department of Statistics, Federal University of São Carlos.
E-mail: gpereira@ufscar.br
- Jianwen Cai, Department of Biostatistics, Gillings School of Global Public Health, University

REFERENCES

of North Carolina at Chapel Hill.

E-mail: cai@bios.unc.edu