University of São Paulo
"Luiz de Queiroz" College of Agriculture

Residuals and diagnostic methods in models for polytomous data

Patricia Peres Araripe

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2022

Patricia Peres Araripe
Degree in Mathematics

Residuals and diagnostic methods in models for polytomous data

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **IDEMAURO ANTONIO RODRIGUES DE LARA**

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2022

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Araripe, Patricia Peres
Residuals and diagnostic methods in models for polytomous data / Patricia Peres Araripe. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2022.
67 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Distribuição multinomial 2. Modelos multicategóricos 3. Análise de resíduos 4. Resíduos quantílicos aleatorizados 5. Distâncias 6. *Half-normal plot* . I. Título.

ACKNOWLEDGMENTS

First, I thank God and Our Lady Aparecida for allowing me to walk this path with faith, hope, and the certainty that everything in my life is for my best in His time.

To my husband, Mateus Urban, for being this partner, and friend, always supporting and encouraging me. He always said, “everything will work out”. I wouldn’t have made it without you by my side. I love you infinitely. Thank you so much. My mother, who helped me so many times, was always present in all struggles, losses, and victories! Thank you, my queen, for everything!

My sisters, Priscila and Tatiana, for all the words of encouragement and moments of joy. And to my nephews who are lightness and happiness in our lives, I love you! To the rest of my family and friends for all the moments lived and those we will still have.

I want to thank my advisor Idemauro who supported and helped me develop this research. He was comprehensible in every moment of my Ph.D. Gratitude! To Rafael and Niamh professors from Maynooth University for all their teaching, support, and patience during this challenging pandemic period. I am immensely grateful that my internship period took place under their co-supervision, professionals that I respect and admire. Without you, this work would not be possible.

Finally, I would like to thank the professors and professionals who work in the Department of Exact Sciences at ESALQ/USP. They are always kind and helpful! To my friends that I made during this Ph.D., Rafaela, Maíra, Pórtya, Vivi, Tati, Janaína, Laura, and Rhayra, who shared special moments with me. I will never forget you, and you live in my heart.

This work was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, incluing the intership period at Maynooth University.

SUMÁRIO

RESUMO

Resíduos e métodos de diagnósticos em modelos para dados politômicos

Experimentos e estudos observacionais que resultam em dados politômicos nominais ou ordinais são conduzidos com frequência em diversas áreas de conhecimento, em especial nas ciências agrárias ou biológicas. O modelo dos logitos generalizados é a alternativa empregada para a análise desse tipo de dados e com base nele obtidas as conclusões e tomadas de decisão. Na inferência estatística, é muito importante validar um modelo que foi ajustado aos dados por meio de métodos de diagnósticos com base em resíduos adequados. No entanto, a análise de resíduos e diagnósticos para modelos associados aos dados politômicos ainda são emergentes na pesquisa científica, constituindo-se em objeto de pesquisa na área de Estatística. Como a variável categórica politômica é multivariada, os resíduos ordinários de Pearson e deviance são vetores por indivíduo com distribuição desconhecida, o que gera desafios na visualização e interpretação gráfica. O resíduo quantílico aleatorizado pode ser utilizado para contornar os problemas com esses resíduos. Entretanto, observa-se que falta uma investigação da sua performance para a regressão politômica por meio de estudos de simulação. Como uma alternativa para reduzir a dimensão dos resíduos e estudar *outliers* este trabalho propõe empregar as medidas de distâncias Euclidiana e de Mahalanobis, uma vez que não se tem registros de sua utilização para o caso multinomial. Nesse contexto, as contribuições metodológicas desse trabalho são: revisão de resíduos existentes para a classe de modelos associados aos dados politômicos; estudo da normalidade dos resíduos quantílicos aleatorizados; proposição do uso das distâncias Euclidiana e de Mahalanobis para reduzir a dimensão dos resíduos ordinários, constituindo-se assim em um procedimento para o diagnóstico dos modelos dos logitos generalizados, permitindo identificar a presença de *outliers*. Duas aplicações ilustram a utilidade do resíduo quantílico aleatorizado e das medidas de distância. A performance dos métodos propostos foram feitas por meio de estudos de simulação. Nesses estudos, avaliou-se o desempenho dos resíduos quantílicos aleatorizados para os dados nominais individuais bem como o uso das distâncias Euclidiana e de Mahalanobis para dados agrupados. Foram empregadas técnicas gráficas como o gráfico meio-normal e o teste Shapiro-Wilk para avaliação da normalidade. Sob diferentes cenários, os estudos de simulação demonstraram que as abordagens são pertinentes para avaliar a bondade do ajuste do modelo dos logitos generalizados aos dados. Adicionalmente, registra-se que tais estudos são apenas o princípio para uma área de pesquisa com muitas lacunas a serem preenchidas.

Palavras-chave: Modelo dos logitos generalizados, Resíduo quantílico aleatorizado, Distâncias, Gráfico meio-normal de probabilidade

ABSTRACT

Residuals and diagnostic methods in models for polytomous data

Experiments and observational studies that result in polytomous data, nominal or ordinal, are frequently conducted in different areas of knowledge, especially in the agricultural or biological sciences. The generalized logit model is the alternative used for the analysis of this type of data and based on it, conclusions and decision-making are obtained. In statistical inference, it is very important to validate a model that has been fitted to the data using diagnostic methods based on appropriate residuals. However, residual analysis and diagnostics for models associated with polytomous response are still emerging in scientific research, constituting an object of research in the area of Statistics. As the polytomous categorical variable is multivariate, Pearson's ordinary residuals and deviance are vectors per individual with unknown distribution, which creates challenges in graphical visualization and interpretation. Randomized quantile residuals can be used to circumvent problems. However, it is observed that there is a lack of an investigation of its performance for the polytomous regression through simulation studies. As an alternative to reduce the dimension of the residuals and study outliers, this work proposes to use Euclidean and Mahalanobis distance measures, since there are no records of their use for the multinomial case. In this context, the methodological contributions of this work are: review of existing residuals for the class of models associated with polytomous data; study of the normality of randomized quantile residuals; proposition of using Euclidean and Mahalanobis distances to reduce the dimension of ordinary residuals, thus constituting a procedure for the diagnosis of generalized logit models, allowing the identification of the presence of outliers. Two applications illustrate the utility of the randomized quantile residuals and distance measurements. The performance of the proposed methods was done through simulation studies. In these studies, we evaluated the performance of randomized quantile residuals for individual nominal data as well as the use of Euclidean and Mahalanobis distances for grouped data. Graphic techniques such as the half-normal plot were used to assess the model and the Shapiro-Wilk test were used to verify normality of residuals. Under different scenarios, simulation studies have shown that the approaches are relevant to assess the goodness of fit of the generalized logits model to the data. Additionally, it is noted that such studies are just the beginning of a research area with many gaps to be filled.

Keywords: Generalized logit models, Randomized quantile residual, Distances, Half-normal plot

1 INTRODUTION

Response variables that represent categories are frequent in scientific research in the agronomic and biological areas. Experiments can be carried out with the interest of studying, for example, the severity of a particular disease in fruits, the level of infestation by pests in a plantation, the classification of plants, and the food preference of insects, among others. The data resulting from these experiments are categorized, that is, discrete data referring to a variable response defined through a finite number of categories (Paulino and Singer, 2006). The categorized variables can be classified according to the number of categories they have, and those that present more than two possible responses are named polytomous. According to Agresti (2002), polytomous variables can be distinguished by two types of measurement scales: nominal (unordered categories) and ordinal (naturally ordered categories).

The statistical models developed for analyzing polytomous data (nominal or ordinal) are based on the multinomial probability distribution, that belongs to the multi-parametric exponential family. These models are an extension of the Generalized Linear Models (GLMs), that was proposed by Nelder and Wedderburn (1972) and described in detail in McCullagh and Nelder (1989). Generalized logit models are commonly used in studies with a categorical response (polytomous), as Agresti (2002) and Tutz (2011) are classic references for these models.

When fitting a model to a data set, it is essential to evaluate possible deviations from the assumptions of this model. The estimates obtained must be resistant to small perturbations both in the model and in the data to not lead the researcher to inferences and inadequate predictions. In this context, residual analysis is one of the most important steps in choosing a statistical model, as it allows to check its assumptions and, consequently, the reliability of the statistical inference based on it (Singer et al., 2017). In the GLM scenarios, the first references to residuals were throughs Pregibon (1981) with a focus on logistic regression models (for two response categories), in addition to Pierce and Schafer (1986) and Williams (1987). However, the extension to the other cases was presented by McCullagh and Nelder (1989). The GLM residuals are used to explore the adequacy of the fitted model concerning the choice of variance function, link function, and terms of the linear predictor. In addition, residuals are essential to indicate the presence of outliers (Cordeiro and Neto, 2004).

In the case of models associated with polytomous categorical response, it is observed that studies employing techniques based on residual analysis to assess the goodness-of-fit of models to the data are still emerging. As the variable is multivariate, since each category corresponds to a dimension of a vector, the ordinary residual, defined as the numerical difference between the observed and fitted values, is also a vector (Reiter and Kohnen, 2005). This multidimensional residual may not be informative when used in diagnostic techniques (formal or informal) for model validation, requiring the development of new residual proposals that began to appear in the literature in the 2000s.

Focusing on the proportional odds model for ordinal data with individual structure, in which an experimental unit is an individual, Liu et al. (2009) defined the cumulative residuals considering a binary response (by grouping the categories) and the vector of cumulative residuals for the original response. As considered in Arbogast and Lin (2005), the authors used the sum of residuals from both approaches to assess the model's fit in relation to the covariates of the linear predictor. However, examining the performance of these residuals on diagnostic plots is not a simple task.

On the other hand, two proposals of one-dimensional residuals can be cited to evaluate the fit of any models that assume proportional odds. The first refers to the residual defined by Li and Shepherd (2012) based on the ordinal variable, also having a discrete nature. The second proposal was presented by Liu and Zhang (2018), who obtained continuous residuals of a continuous variable defined to replace the original one. Furthermore, Liu and Zhang (2018) showed that the residuals in their approach had expected patterns in diagnostic plots, unlike the residuals of Li and Shepherd (2012), which displayed

unusual patterns in the different scenarios in which the model was correctly specified to the data.

For the context of individual nominal data, Cheng et al. (2021) defined a residual vector to evaluate the fit of discrete models, considering the methodology developed by Liu and Zhang (2018). As the vector has a continuous multidimensional distribution, the values obtained in each dimension were evaluated in several plots and diagnostic tests. The authors showed good results to detect nonlinear covariate and interaction effects for the generalized logit model associated with data with three categories. However, the bigger the number of categories of the variable response, the greater the dimension of the residual vector, and the number of values can make it difficult to interpret the behavior of residuals in plots and diagnostic tests.

The polytomous data can still be in the grouped structure, in which an experimental unit is a group of individuals, and these are generally arranged in contingency tables. Andersen (1992) used the leverage measure and Cook's distance to measure the influence on the estimates of the parameters of the RC association model (Goodman, 1985) when deleting all observations of a specific cell in the contingency table. In analyzing grouped data in a longitudinal study of toxicological mortality, O'Hara Hines et al. (1992) applied the measure of local influence to evaluate the effect of small perturbations on the cumulative model assumptions. Seber and Nyangoma (2000) defined a vector of residuals, so-called projected residuals, to evaluate log-linear models based on a more complex approach introduced by Cook and Tsai (1985) in nonlinear theory. The elements of the vector must be approximately distributed by the standard normal and present a small magnitude referring to a bias term. In the work of Silva (2003), the residuals were defined to evaluate the generalized logit model with three response categories. As the model is composed of two equations in terms of logits with different parameters, the author presented for each of the sub-models the standardized Pearson and deviance residuals, without sign assignment, using them in diagnostic plots. Furthermore, Gupta et al. (2008) presented Pearson's residual vector to evaluate the generalized logit models, with parameters estimated using the minimum phi-divergence estimator instead of maximum likelihood.

In short, despite the contributions in the development of residuals to evaluate this class of models, there is still a need to stimulate new methodologies that can help researchers in this important area. In this context, the main goal and the specific contributions of this present thesis are given below.

1.1 Main goal

The purposes of this work are:

i) To present a review about diagnostic techniques based on the residual analysis for polytomous categorical data.

ii) To develop diagnostic techniques (formal and informal) based on the residuals analysis for the models, in which response variable is categorical and polytomous, in particular, nominal nature.

1.2 Specific contributions

The specific contributions of this thesis are:

i) The using the randomized quantile residuals related to the generalized logit model and the demonstration that they approximately follow a normal distribution when correctly specified for individual nominal data, as well as to show the power of the Shapiro-Wilk test for this case by simulation studies.

ii) The application the Euclidean and Mahalanobis distances to reduce the dimension of the ordinary residuals for the generalized logit models, as well as the demonstration that these measures, when

correctly specified for grouped nominal data, can detect outliers using the half-normal plot with a simulated envelope.

1.3 Organization of thesis

This thesis is organized as follows: this first chapter describes the introduction, with a literature review of the theoretical framework, objectives and contribution of the work to the scientific area. The second chapter reviews the residuals associated with the ordinal response, specially the surrogate residual proposed by Liu and Zhang (2018) for the proportional odds model. As an illustration of this residual analysis, it was presented a field study with Tambaqui fish, example from literature, to verify the relationship between the different types of genotypes in the classification of the lesion found in the liver, in which the response variable is ordinal.

In the third chapter, two proposals are used to verify the assumptions of the generalized logit model associated with nominal data under simulation studies. For individual nominal data, the performance and distribution of randomized quantile residuals were examined in diagnostic plots and the normality assessed by the Shapiro-Wilk test. Furthermore, Euclidean and Mahalanobis distances were proposed to reduce the dimension of the ordinary residuals for the case of grouped data. These measures were examined through the half-normal plot with a simulated envelope to detect outliers. Two applications are presented to illustrate the perform of the proposed diagnostic techniques. Finally, in the fourth chapter, it is done the final considerations.

References

Agresti, A. (2002). *An introduction to categorical data analysis*. John Wiley & Sons, Nova Jersey, 3 edition.

Andersen, E. B. (1992). Diagnostics in categorical data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):781–791.

Arbogast, P. G. and Lin, D. (2005). Model-checking techniques for stratified case-control studies. *Statistics in medicine*, 24(2):229–247.

Cheng, C., Wang, R., and Zhang, H. (2021). Surrogate residuals for discrete choice models. *Journal of Computational and Graphical Statistics*, 30(1):67–77.

Cook, R. D. and Tsai, C. L. (1985). Residuals in nonlinear regression. *Biometrika*, 72(1):23–29.

Cordeiro, G. M. and Neto, E. A. L. (2004). *Modelos paramétricos*.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, pages 10–69.

Gupta, A. K., Nguyen, T., and Pardo, L. (2008). Residuals for polytomous logistic regression models based on φ-divergences test statistics. *Statistics*, 42(6):495–514.

Li, C. and Shepherd, B. E. (2012). A new residual for ordinal outcomes. *Biometrika*, 99(2):473–480.

Liu, D. and Zhang, H. (2018). Residuals and diagnostics for ordinal regression models: A surrogate approach. *Journal of the American Statistical Association*, 113(522):845–854.

Liu, I., Mukherjee, B., Suesse, T., Sparrow, D., and Park, S. K. (2009). Graphical diagnostics to check model misspecification for the proportional odds regression model. *Statistics in medicine*, 28(3):412–429.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2 edition.

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

O'Hara Hines, R. J., Lawless, J. F., and Carter, E. M. (1992). Diagnostics for a cumulative multinomial generalized linear model, with applications to grouped toxicological mortality data. *Journal of the American Statistical Association*, 87(420):1059–1069.

Paulino, C. D. M. and Singer, J. M. (2006). *Análise de dados categorizados*. Edgard Blücher, São Paulo.

Pierce, D. A. and Schafer, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, 81(396):977–986.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of statistics*, 9(4):705–724.

Reiter, J. P. and Kohnen, C. N. (2005). Categorical data regression diagnostics for remote access servers. *Journal of Statistical Computation and Simulation*, 75(11):889–903.

Seber, G. and Nyangoma, S. (2000). Residuals for multinomial models. *Biometrika*, 87(1):183–191.

Silva, J. A. P. (2003). *Métodos de diagnóstico em modelos logísticos trinomiais*. Dissertação (mestrado em estatística).

Singer, J. M., Rocha, F. M. M., and Nobre, J. S. (2017). Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review*, 85(2):290–324.

Tutz, G. (2011). *Regression for categorical data*, volume 34. Cambridge University Press, Cambridge.

Williams, D. A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(2):181–191.

2 ORDINAL DATA AND RESIDUAL ANALYSIS: REVIEW AND APPLICATION

Abstract

Experiments in which the response is ordinal polytomous are often performed in the agricultural sciences and, often, the cumulative logit models are used to analyze this variable. A particular characteristic is that the polytomous variables are objects of multivariate statistics and the ordinary residual, associated with the classical models available, is a vector for each individual. Consequently, these residuals are not easily interpreted, and their distribution is unknown. Residual analysis is an essential step in validating any statistical model, and not performing it can allow a model to incorrectly fit the data, resulting in erroneous conclusions and inferences. In this context, the work aims to review the residuals for ordinal data available in the literature, emphasizing the so-called surrogate residuals with continuous distribution. As a practical application, it is present an experiment carried out with Tambaqui fish of different types of genotype. The response variable in this study is the severity of the lesions found in the livers of Tambaquis. The estimation of the parameters was performed using the maximum likelihood. The selected model by the likelihood ratio test included the proportional odds and fish genotype effect. According to this model, it was possible to verify in this study that fish with genotype 122 presented a higher probability of liver lesion classified as irreversible $(71,26\%)$, while Tambaquis with genotype 130 had a higher probability of moderate lesion, $46,75\%$. For the model diagnostics, the half-normal plot and the Kolmogorov-Smirnov test were used to examine the performance of the surrogate residual. The results obtained provided evidence of the adequacy of the selected model since the residuals did not reveal patterns or influential points in diagnostic tools.

Keywords: Cumulative logit model; Maximum likelihood; Half-normal plot; Kolmogorov-Smirnov test.

2.1 Introdution

In agricultural sciences, it is common to carry out experiments that result in polytomous data as a response of interest. These data assume values in a finite set of categories with nominal or ordinal scale (natural ordering between categories) and have a multinomial distribution regardless of this nature (Agresti, 2002). The models with the logit link function are the most used in the statistical analysis of these data. The proportional odds model (McCullagh, 1980) is widely used for the ordinal case with a smaller number of parameters due to the assumption of proportionality (Tutz, 2011). However, other alternatives can be considered, such as the cumulative probit model or the Proportional Hazards model with a complementary log-log link function (Agresti, 2010b). When the proportionality assumption is not valid, the cumulative logit model (Williams and Grizzle, 1972) can be fitted to the data or the adjacent-categories logit model, for example (Ananth and Kleinbaum, 1997 and Agresti, 2002). Furthermore, one can assume another discrete multivariate distribution for the response variable, such as the Dirichlet distribution, which is the conjugate distribution of the multinomial in Bayesian inference (Ng et al., 2011).

When selecting a model, it is essential to assess the quality of its fit to the data as well as to validate its assumptions. The fitted model must describe the observed data well so as not to result in incorrect inferences. In this context, residual analysis plays an important role in detecting possible failures resulting from the fit and identifying outliers and/or influential points, becoming an integral part of any regression problem (Cook and Weisberg, 1982). McCullagh and Nelder (1989) paid substantial attention to defining residuals for Generalized Linear Models (GLMs), with Pearson and deviance residuals frequently used in the diagnostics of GLMs. However, these residuals do not apply to multinomial data due to the nature of the response variable. As the polytomous variable is multivariate, the ordinary residual given by the difference between the observed response and the estimated probability is a vector

for each individual (Reiter and Kohnen, 2005). Therefore, diagnostic plots of residuals are difficult to interpret since their distribution is difficult to identify. Furthermore, few papers in the literature involve types of residuals that help validate models associated with polytomous data, and these are defined, in particular, for the case in which the response of individual results in only one of the categories.

For the ordinal case, Liu et al. (2009) presented the vector of cumulative residuals focusing on validating the proportional odds model with respect to the covariates of the linear predictor. However, it is not simple to interpret the behavior of these residuals in diagnostic plots, as is the case with residuals for continuous variables. Li and Shepherd (2012) and Liu and Zhang (2018) defined residuals that correspond to a single value per individual regardless of the number of categories. While the residual proposed by Li and Shepherd (2012) is obtained in the discrete space of the original response, in the approach used Liu and Zhang (2018), a continuous variable replaces the ordinal variable, and the residual is defined through this new variable. Liu and Zhang (2018) compared the performance of the residuals so-called surrogate, with those proposed by Li and Shepherd (2012) in the residuals versus covariates plot and Quantile-Quantile plot (Q-Q plot) to assess the fit of the cumulative probit model with respect to mean structure, heteroscedasticity, and proportionality. The authors showed that the surrogate residuals presented expected behaviors in these plots for the model correctly specified to the data. In contrast, the residuals defined by Li and Shepherd (2012) showed unusual patterns that did not allow concluding in favor of the correct model.

The aim of this work is to present a review of models and residuals for polytomous ordinal data, considering the relevance and need for studies and research in this area. As a specific case, we show the performance of the surrogate residuals to evaluate the cumulative logit model for ordinal response. As a motivational study and application, it is presented the research carried out with Tambaqui fish (*Colossoma macropomum*), in which a type of histopathological alteration was observed in the liver fish. Therefore, in this study, the response variable is the severity of lesion found in the fish liver (natural ordering), which was classified as mild, moderate, and irreversible. Also, it is verified the relationship of the classifications with the different gene expressions of the Tambaquis. This species is a source of aquatic protein widely consumed in the North region of Brazil and has attracted significant interest from fish farmers from other countries (Lopes et al., 2016). Given the large production of Tambaqui in the country, the aquatic environment and the management of these fish must be appropriately controlled to generate a healthy population, not causing losses in productivity (Correa et al., 2018).

2.2 Models for ordinal response

When the response variable Y_i takes on a value in the set $\{1, 2, \ldots, J\}$ for the i-th individual, $i = 1, 2, \ldots, n$, with the ordered categories $1 < 2 < \ldots < J$ and multinomial distribution, the cumulative logit models with canonical the link function can be used to describe the functional relationship between the response and covariates of the study. According to Agresti (2010b), models that consider the natural order of the response can produce more powerful results than models that ignore ordinality.

2.2.1 Cumulative logit model

The cumulative logit model (Williams and Grizzle, 1972) is a multivariate extension in the class of generalized linear models used to model the dependence of an ordinal response on discrete or continuous covariates. This model is defined by:

$$\text{logit}\left[\gamma_{ij}(\mathbf{x}_i)\right] = \log\left[\frac{\gamma_{ij}(\mathbf{x}_i)}{1-\gamma_{ij}(\mathbf{x}_i)}\right] = \alpha_j + \sum_{k=1}^{p}\beta_{jk}x_{ik} = \alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i,\ j = 1, \ldots, J-1, \tag{2.1}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is the vector of the p covariates for the i-th individual, $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})'$ represents the vector of regression parameters and α_j is the intercept, with $j = 1, 2, \ldots, J-1$. Here, $\gamma_{ij}(\mathbf{x}_i)$ is the cumulative probability of the individual i until the j-th category, that is, $\gamma_{ij}(\mathbf{x}_i) = P(Y_i \leq j|\mathbf{x}_i) = \pi_{i1}(\mathbf{x}_i) + \ldots + \pi_{ij}(\mathbf{x}_i)$, $j = 1, \ldots, J$, being $\pi_{ij}(\mathbf{x}_i) = P(Y_i \leq j|\mathbf{x}_i) - P(Y_i \leq j-1|\mathbf{x}_i)$ the probability of the (marginal) response in the j-th category, more precisely,

$$\pi_{ij}(\mathbf{x}_i) = \frac{\exp(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}_{j-1}'\mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}_{j-1}'\mathbf{x}_i)}$$

with $P(Y_i \leq 0|\mathbf{x}_i) = 0$ and $P(Y_i \leq J|\mathbf{x}_i) = 1$.

In the cumulative logit model, the regression parameters are not constant for the j logits, i.e., $\boldsymbol{\beta}_j$ can vary according to each response category. The estimation of the parameters of the model (2.1) is generally performed using the maximum likelihood method, whose likelihood function for the random sample of size n is given by

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} [\pi_{ij}(\mathbf{x}_i)]^{y_{ij}} \right\} \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} [P(Y_i \leq j|\mathbf{x}_i) - P(Y_i \leq j-1|\mathbf{x}_i)]^{y_{ij}} \right\} \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} \left[\frac{\exp(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}_{j-1}'\mathbf{x}_i)}{1+\exp(\alpha_{j-1} + \boldsymbol{\beta}_{j-1}'\mathbf{x}_i)} \right]^{y_{ij}} \right\}, \end{aligned}$$

where $y_{ij} = 1$ if the response of individual i, $i = 1, \ldots, n$, belongs to the category j, $j = 1, \ldots, J$, $y_{ij} = 0$ otherwise, with $\sum_{j=1}^{J} y_{ij} = 1$ and $\boldsymbol{\theta} = \left(\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{J-1}\right)'$ is the vector with the parameters to be estimated. It is necessary to use iterative methods such as the Newton-Raphson method to maximize L and obtain the maximum likelihood estimators of the parameters (Agresti, 2002).

An alternative to model (2.1) is the proportional odds model, which assumes that the effects of the covariates are the same for each logit j, resulting in a more parsimonious model, that is, with a smaller number of parameters (Bilder and Loughin, 2014). The proportional odds assumption results in the simplest fit with easy interpretation, but it should always be carefully verified (Lemos et al., 2015).

2.2.2 Proportional odds model

The simplest model in the class of cumulative logit models involves parallel regressions on the ordinal scale and assumes equivalent proportions by assuming the same regression parameter for all categories. This model, called the proportional odds model, was introduced by McCullagh (1980) and can be expressed by

$$\text{logit}\left[\gamma_{ij}(\mathbf{x}_i)\right] = \log\left[\frac{\gamma_{ij}(\mathbf{x}_i)}{1-\gamma_{ij}(\mathbf{x}_i)}\right] = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} = \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i, \quad j = 1, \ldots, J-1, \tag{2.2}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is the vector of covariates for the individual i, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ represents the vector of regression parameters, α_j is the intercept, and the last category J as the reference. Here, $\gamma_{ij}(\mathbf{x}_i)$ is the cumulative probability of individual i until the j-th category, that is, $\gamma_{ij}(\mathbf{x}_i) = P(Y_i \leq j|\mathbf{x}_i) = \pi_{i1}(\mathbf{x}_i) + \ldots + \pi_{ij}(\mathbf{x}_i)$, $j = 1, \ldots, J$. The probabilities $\pi_{ij}(\mathbf{x}_i)$ are obtained for the model (2.2) by means of subtractions given by

$$\pi_{ij}(\mathbf{x}_i) = P(Y_i \leq j|\mathbf{x}_i) - P(Y_i \leq j-1|\mathbf{x}_i) = \frac{\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)}{1 + \exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}{1 + \exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)},$$

where $P(Y_i \leq 0|\mathbf{x}_i) = 0$ and $P(Y_i \leq J|\mathbf{x}_i) = 1$.

As the effects of the covariates are equal, the model assumes that the effects on the logit are identical for all categories of the response variable. Then, the $J-1$ logits are shifted only as a function of the intercept (Bilder and Loughin, 2014). According to Agresti (2007), the maximum likelihood procedure can be used to estimate the parameters of the model (2.2), with a likelihood function for the random sample of dimension n described by

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} [\pi_{ij}(\mathbf{x}_i)]^{y_{ij}} \right\} \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} [P(Y_i \leq j|\mathbf{x}_i) - P(Y_i \leq j-1|\mathbf{x}_i)]^{y_{ij}} \right\} \\ &= \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} \left[\frac{\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i)} \right]^{y_{ij}} \right\}, \end{aligned}$$

where $y_{ij} = 1$ if the response of individual i, $i = 1, \ldots, n$, belongs to category j e $y_{ij} = 0$ otherwise, $j = 1, \ldots, J$, with $\sum_{j=1}^{J} y_{ij} = 1$ and $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\beta})'$ representing the vector of parameters. According to McCullagh (1980), the Newton-Raphson method with Fisher scoring can be used to obtain parameter estimates, converging rapidly even with poor initial values.

The odds ratio is a very intuitive and used way to interpret the parameters estimated by the proportional odds model. Consider two subpopulations characterized by vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, then the cumulative odds ratio for the two subpopulations is given by

$$\frac{P(Y_i \leq j|\mathbf{x}_1)/P(Y_i > j|\mathbf{x}_1)}{P(Y_i \leq j|\mathbf{x}_2)/P(Y_i > j|\mathbf{x}_2)} = \exp\left[\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)\right], \ j = 1, 2, \ldots, J-1,$$

where the odds of occurring $\{Y_i \leq j|\mathbf{x}_i = \mathbf{x}_1\}$ is equal to $\exp[\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)]$ times the odds of occurring $\{Y_i \leq j|\mathbf{x}_i = \mathbf{x}_2\}$. As stated in Bilder and Loughin (2014), the cumulative odds ratio remains the same regardless of the category j used, and this is due to the assumption that the effects of the covariates are the same for all categories.

As the proportional odds model is a particular case of the model (2.1), the proportionality assumption can be verified through the likelihood ratio test (LRT) with the following hypotheses

$$\begin{cases} H_0 : \boldsymbol{\beta}'_j = \boldsymbol{\beta}', \ \forall j = 1, 2, \ldots, J-1 \\ H_1 : \boldsymbol{\beta}'_j \neq \boldsymbol{\beta}', \ \exists \text{ at least a } j \end{cases}$$

and with the statistic of the test given by

$$\Lambda = -2\log\left[\frac{L_{H_0}}{L_{H_1}}\right] \sim \chi^2_m,$$

where L_{H_0} is the likelihood function under the null hypothesis H_0, i.e., referring to model (2.2) and L_{H_1} is the likelihood function under the alternative hypothesis H_1, i.e., referring to model (2.1). Here, Λ follows an approximate Chi-square distribution, in which the degrees of freedom, m, are obtained by the difference between the numbers of the parameters under the hypotheses H_0 and H_1. If the null hypothesis is not rejected at the 5% significance level, then the proportional odds model can be fitted to the data (Lemos et al., 2015 and Giolo, 2017).

The proportionality assumption can be verified in two ways: global and individual. Globally, all model covariates are considered, while individually, it is considered covariate by covariate. In the case of rejection of the null hypothesis for part of the covariates, that is, some covariates have the property of proportional odds and others do not, an alternative is the partial proportional odds model (Agresti, 2010b).

2.2.3 Partial proportional odds model

The proportional odds assumption is not always achieved in practice. A model proposed by Peterson and Harrell Jr (1990), an extension of the proportional odds model, can be used when part of the covariates violates this assumption.

Consider the vector $\mathbf{x}_i$ with the values of p covariates for the i-th individual that present proportional odds and the vector $\mathbf{z}_i$ with the values of q ($q \leq p$) covariates that do not, so the partial proportional odds model is given by

$$\text{logit}\left[\gamma_{ij}(\mathbf{x}_i, \mathbf{z}_i)\right] = \log\left[\frac{\gamma_{ij}(\mathbf{x}_i, \mathbf{z}_i)}{1-\gamma_{ij}(\mathbf{x}_i, \mathbf{z}_i)}\right] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\varrho}_j'\mathbf{z}_i, \quad j = 1, \ldots, J-1, \tag{2.3}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ and $\boldsymbol{\varrho}_j = (\varrho_{j1}, \varrho_{j2}, \ldots, \varrho_{jq})'$ sare the vectors of regression parameters, α_j is the intercept and the last category taken as a reference. Here, the vector $\boldsymbol{\varrho}_j$ describes the effect of non-proportionality for each $j-$th cumulative logit associated with the vector $\mathbf{z}_i$. In this model, $J-1$ intercepts, p coefficients referring to the vector $\boldsymbol{\beta}$, which are independent of the compared categories, and $q(J-1)$ coefficients referring to the vector $\boldsymbol{\varrho}_j$ are estimated. Furthermore, $\gamma_{ij}(\mathbf{x}_i, \mathbf{z}_i)$ is the cumulative probability of individual i until the j-th category, i.e., $P(Y_i \leq j|\mathbf{x}_i, \mathbf{z}_i) = \pi_{i1}(\mathbf{x}_i, \mathbf{z}_i) + \ldots + \pi_{ij}(\mathbf{x}_i, \mathbf{z}_i)$, $j = 1, \ldots, J$, and the probabilities $\pi_{ij}(\mathbf{x}_i, \mathbf{z}_i)$ for th model (2.3) are obtained in an analogous way to those obtained for models (2.1) and (2.2), so

$$\begin{aligned}\pi_{ij}(\mathbf{x}_i, \mathbf{z}_i) &= P(Y_i \leq j|\mathbf{x}_i, \mathbf{z}_i) - P(Y_i \leq j-1|\mathbf{x}_i, \mathbf{z}_i) \\ &= \frac{\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\varrho}_j'\mathbf{z}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\varrho}_j'\mathbf{z}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\varrho}_j'\mathbf{z}_i)}{1+\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\varrho}_j'\mathbf{z}_i)},\end{aligned}$$

where $P(Y_i \leq 0|\mathbf{x}_i, \mathbf{z}_i) = 0$ and $P(Y_i \leq J|\mathbf{x}_i, \mathbf{z}_i) = 1$.

The estimation of parameters can also be performed using the maximum likelihood method for the random sample of size n (Agresti, 2010b). Considering $y_{ij} = 1$ if the response of individual i, $i = 1, \ldots, n$, belongs the category j, $j = 1, \ldots, J$, $y_{ij} = 0$ otherwise and $\sum_{i=1}^{J} y_{ij} = 1$, the estimators of the model (2.3) can be obtained by maximizing the likelihood function (or its logarithm) given by

$$\begin{aligned}L(\boldsymbol{\theta}) &= \prod_{i=1}^{n}\left\{\prod_{j=1}^{J}\left[\pi_{ij}(\mathbf{x}_i, \mathbf{z}_i)\right]^{y_{ij}}\right\} \\ &= \prod_{i=1}^{n}\left\{\prod_{j=1}^{J}\left[P(Y_i \leq j|\mathbf{x}_i, \mathbf{z}_i) - P(Y_i \leq j-1|\mathbf{x}_i, \mathbf{z}_i)\right]^{y_{ij}}\right\} \\ &= \prod_{i=1}^{n}\left\{\prod_{j=1}^{J}\left[\frac{\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\varrho}_j'\mathbf{z}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\gamma}_j'\mathbf{z}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\varrho}_j'\mathbf{z}_i)}{1+\exp(\alpha_{j-1} + \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\varrho}_j'\mathbf{z}_i)}\right]^{y_{ij}}\right\},\end{aligned}$$

where $\boldsymbol{\theta} = \left(\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\beta}, \boldsymbol{\varrho}_1, \ldots, \boldsymbol{\varrho}_{J-1}\right)'$ corresponds to the vector of parameters to be estimated. The estimates can be obtained using the step-halving technique in the modified Gauss-Newton algorithm that ensure, in each iteration, an increase in the likelihood logarithm (Peterson and Harrell Jr, 1990).

The adjacent-categories logit model is also an alternative when the proportionality assumption is not satisfied. It considers the ratio between the probabilities of successive categories rather than the cumulative probabilities. Additionally, it is possible to find this model and others for ordinal data in Ananth and Kleinbaum (1997), Agresti (2002), Agresti (2007), Agresti (2010b), Tutz (2011), Bilder and Loughin (2014), Giolo (2017), among others.

2.3 Residuals for ordinal data

After fitting a model to the data, it is essential to verify whether its assumptions are satisfied and identify individuals that may disproportionately interfere with the results obtained. Through an

analysis of the residuals, it is possible to study the robustness of the fitted model in terms of the various aspects that involve its formulation and the estimates of its parameters, detecting potential problems, and improving the fitting process (Souza, 2006).

2.3.1 Ordinary Residual

For the class of models with a polytomous categorical response, the ordinary residual associated with the i-th individual, $i = 1, \ldots, n$, is a vector $J \times 1$ defined by (Reiter and Kohnen, 2005)

$$\hat{\mathbf{r}}_i = \mathbf{y}_i - \hat{\boldsymbol{\pi}}_i = (y_{i1} - \hat{\pi}_{i1}, y_{i2} - \hat{\pi}_{i2}, \ldots, y_{iJ} - \hat{\pi}_{iJ})', \tag{2.4}$$

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})'$ is the observed vector with $y_{ij} = 1$ if the response of the individual i belongs to the category j and $y_{ij} = 0$ otherwise, $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i1}, \hat{\pi}_{i2}, \ldots, \hat{\pi}_{iJ})'$ is the estimated probabilities vector. The only positive element in this vector pertains to the observed outcome for the individual. This vector may not be informative in the diagnostic techniques for analyzing residuals since its asymptotic distribution is unknown.

2.3.2 Cumulative Residual

Specifically for the proportional odds model, defined in section 2.2.2, Liu et al. (2009) presented the cumulative residuals for a binary response (by collapsing the categories) and the vector of cumulative residuals considering the original response. For the multivariate case, the vector of cumulative residuals, $J \times 1$, for each individual is expressed by

$$\mathbf{r}_i^* = \mathbf{y}_i - \boldsymbol{\gamma}_i = (y_{i1} - P(Y_i \leq 1|\mathbf{x}_i), y_{i2} - P(Y_i \leq 2|\mathbf{x}_i), \ldots, y_{iJ} - P(Y_i \leq J|\mathbf{x}_i))',$$

where $\boldsymbol{\gamma}_i = (P(Y_i \leq 1|\mathbf{x}_i), P(Y_i \leq 2|\mathbf{x}_i), \ldots, P(Y_i \leq J|\mathbf{x}_i))'$ is the vector of cumulative probabilities for the i-th individual. The authors used the sum of this residual vector in graphical and numerical methods to assess the goodness-of-fit of the model. The methods generalize those developed by Arbogast and Lin (2005) for the logistic regression model with binary responses. However, diagnostic plots associated with residuals are difficult to interpret.

2.3.3 LS Residual

Considering the models that assume the assumption of proportionality for the regression parameters, Li and Shepherd (2012) proposed a residual that is a single value per individual, regardless of the number of ordered categories. This residual, called LS, is obtained by the difference between two cumulative probabilities, and the authors examined several properties to apply it to the available diagnostic tools. The residual associated with an individual considering the model 2.2 is obtained by

$$\begin{aligned} R_i^{LS} &= P(Y_i < j|\mathbf{x}_i) - P(Y_i > j|\mathbf{x}_i) \\ &= P(Y_i \leq j-1|\mathbf{x}_i) - [1 - P(Y_i \leq j|\mathbf{x}_i)] \\ &= P(Y_i \leq j-1|\mathbf{x}_i) + P(Y_i \leq j|\mathbf{x}_i) - 1, \end{aligned}$$

with its value varying in the numeric interval of $[-1, 1]$. The Q-Q plot of this residual is obtained compared to the theoretical quantiles of a Uniform distribution in $[-1, 1]$. However, the residual is defined on the discrete space of the response variable, and its conditional distribution can vary according to the covariates. These facts make it difficult to analyze the residuals in diagnostic plots since they do not produce the expected patterns. According to Liu and Zhang (2018), the use of this residual is limited to verifying its zero mean under the correct model.

2.3.4 Surrogate Residual

The residual defined by Liu and Zhang (2018) is also a single value per individual for the models that assumes the proportional odds. Consider the model (2.2) and a latent variable given by $Z_i = -\boldsymbol{\beta}'\mathbf{x}_i + \varepsilon_i$, $i = 1, 2, \ldots, n$, where $\varepsilon_1, \ldots, \varepsilon_n$ is a random sample of the variable ε which follows a standard logistic distribution, $\varepsilon \sim \log(0, 1)$, with probability density function and cumulative distribution function, respectively, given by

$$g(u) = \frac{e^{-u}}{(1+e^{-u})^2} \quad \text{e} \quad G(u) = \frac{e^{u}}{1+e^{u}},$$

where $u \in \mathbb{R}$. The mean and variance of ε are $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \frac{\pi^2}{3}$, respectively.

The concept of latent variable induces a joint distribution of the variables Y_i and Z_i determined by $Y_i = j$ if $\alpha_{j-1} < Z_i \leq \alpha_j$, $j = 1, 2, \ldots, J$, with $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_{J-1} < \alpha_J = \infty$. Thus, the marginal distribution of the ordinal variable Y_i is the same as the distribution specified by the model (2.2). The authors defined a continuous variable S_i based on the conditional distribution of Z_i given Y_i, i.e., S_i follows a truncated distribution of Z_i in the interval $(\alpha_{j-1}; \alpha_j)$ given $Y_i = j$. Therefore, the surrogate residual is obtained by the difference between the surrogate variable and its expected value, with the expression given by

$$R_i^S = S_i - \mathrm{E}_0(S_i|\mathbf{x}_i) = S_i - \mathrm{E}(Z_i|\mathbf{x}_i) = S_i + \boldsymbol{\beta}'\mathbf{x}_i - \int_{-\infty}^{+\infty} u dG(u) \tag{2.5}$$

where E(.) denotes the mean. If the model (2.1) is specified correctly, the variable S_i follows the same distribution of Z_i and the residual R_i^S, which is also a continuous variable, has the following properties:

i) $\mathrm{E}\left(R_i^S|\mathbf{x}_i\right) = 0$;

ii) $\mathrm{Var}\left(R_i^S|\mathbf{x}_i\right) = \frac{\pi^2}{3}$, a constant does not depend on $\mathbf{x}_i$;

iii) Reference distribution: Independent of $\mathbf{x}_i$, the empirical distribution of R_i^S approximates of the standard logistic distribution, that is, $R_i^S \sim G(.)$.

These properties allow an analysis of residuals in practically all existing diagnostic tools for continuous variables (Liu and Zhang, 2018). As the residuals are obtained by random sampling, diagnostic plots may vary from one sample to another (especially for small samples). The authors presented a bootstrap algorithm for the residual (2.5) similar to the bootstrap algorithm used in linear regression proposed by Efron (1979) to account for the variability of conditional sampling. It consists of repeatedly resampling the observed data, generating new data sets, and finding characteristics of interest in the population studied.

The algorithm for obtaining the b-th bootstrap replication of surrogate residuals, $b = 1, 2, \ldots, B$, is given in two steps (Liu and Zhang, 2018):

1) Generate a bootstrap sample of size n through sampling with replacement of the original data and the corresponding covariates, i.e., $\{(\mathbf{x}_{1b}^*, Y_{1b}^*), (\mathbf{x}_{2b}^*, Y_{2b}^*), \ldots, (\mathbf{x}_{nb}^*, Y_{nb}^*)\}$.

2) Using the bootstrap sample obtained in step 1, perform the conditional sampling procedure presented in this section to generate a sample of the surrogate residuals given by $R_{1b}^{S^*}, R_{2b}^{S^*}, \ldots R_{nb}^{S^*}$.

Thus, it is possible to examine the discrepancy between the empirical bootstrap distributions and the reference distribution (standard logistic). As the bootstrap samples are drawn independently, the behavior of $B \times n$ surrogate residuals is examined in the plot of residuals versus covariate (or fitted values), while the median of the B bootstrap distributions is examined in the Q-Q plot.

2.4 Diagnostic techniques

Several diagnostic techniques based on residual analysis can assess the goodness-of-fit of a statistical model. These can be informal through residual plots or formal when using tests. The tests provide a p-value referring to a tested hypothesis. At the same time, the graphical representation is an important exploratory diagnostic feature that can reveal which components of the model were not correctly specified.

When fitting a linear regression model, the Shapiro-Wilk test (Shapiro and Wilk, 1965) is generally used to verify the normality assumption of residuals. On the other hand, the Kolmogorov-Smirnov test (Kolmogorov, 1933) is a widely known test that considers continuous models other than the linear regression model. Through this test, it is possible to examine the degree of agreement between the empirical distribution function of the residuals concerning the theoretical distribution function of reference (Dufour et al., 1998). In addition, a simple way to visualize the shape of the residual distribution is through a histogram, making it possible to compare the result obtained with the shape of the normal distribution or any other distribution.

Consider $R_1, R_2, \ldots, R_n$ a random sample of residuals with empirical distribution function $Q_n(c; R_1, R_2, \ldots, R_n)$ and $G(c)$ the theoretical distribution function of reference. The hypotheses of the Kolmogorov-Smirnov test are given by

$$\begin{cases} H_0 : Q_n(c; R_1, R_2, \ldots, R_n) = G(c), \ \ \forall c \in (-\infty; +\infty) \\ H_1 : Q_n(c; R_1, R_2, \ldots, R_n) \neq G(c), \ \ \exists \text{ at least a c} \end{cases}$$

and test statistic

$$T_n(R_1, R_2, \ldots, R_n) \equiv n^{1/2} d_{KS}\left(Q_n, G\right),$$

where $d_{KS}\left(Q_n, G\right) = \sup_{c \in \mathbb{R}} |Q_n(c; R_1, R_2, \ldots, R_n) - G(c)|$ corresponds to the largest vertical difference between the two distribution functions. For a significance level $\alpha = 5\%$, the H_0 is rejected if the statistic T_n exceeds the quantile value of $1 - \alpha$ as given by the table of quantiles for the Kolmogorov test statistic. In case of non-rejection of the null hypothesis, $R_1, R_2, \ldots, R_n$ is a random sample from the theoretical distribution function.

Although goodness-of-fit tests provide a p-value that indicates how strong the evidence (observed data) is against the null hypothesis, they may fail in certain circumstances, for example, when the sample size is small. Generally, graphical techniques can be more informative, providing a better diagnostics of model adequacy than hypothesis testing (Moral et al., 2017). Among the different types of diagnostic plots, some principals are (Paula, 2013; Faraway, 2016; Moral et al., 2017; among others):

i) Residuals versus covariates: indicates whether the systematic part was incorrectly specified, with the need to include higher-order terms or transform the quantitative covariates into the linear predictor. The expected pattern of this plot is a zero-centered distribution of residuals with constant amplitude;

ii) Residuals versus fitted values: the behavior of the residuals in this plot must be the same as described in item (i) for a well-fitted model. This plot can reveal the existence of heterogeneity of variance in addition to outliers;

iii) Normal and half-normal plots: they are widely used for the diagnostics of the model, being possible to detect outliers and identify failures in the specification of the link function or distribution of the random component. The residuals should follow approximately a straight line with a slope of 45^{o} for a well-fitted model.

Under the normality assumption, the normal plot of the residuals against the expected sorted values of the standard normal distribution, which is approximated by

$$\Phi^{-1}\left[\frac{(i-3/8)}{n+1/4}\right],$$

while in the half-normal plot, the absolute values of the residuals (even with unknown distribution) are compared concerning the expected order statistics of the half-normal distribution, obtained by

$$\Phi^{-1}\left[\frac{(i+n-1/8)}{2n+1/2}\right],$$

where Φ^{-1} is the standard normal distribution function, with $i = 1, \ldots, n$ and n corresponding to the sample size. However, the interpretation of the behavior of the points in these plots can be subjective, and it is difficult to point out other causes for unavoidable irregularities. To assist in visual analysis, Atkinson (1985) proposed adding a simulated envelope to these plots. So, it is possible to observe the proportion of points within the envelope and decide whether the observed residuals are consistent with the fitted model. Generally, there is evidence of a good fit when the number of points outside the envelope equals or less than 5%.

In addition, it is important to examine the existence of one or more points poorly fitted by the model (do not follow the same pattern as the others) and may cause a significant impact on some characteristics of interest, such as the parameter estimate or the corresponding standard error (Singer et al., 2017). A simple technique introduced by Cook (Cook, 1977) that can be used is the deletion, which measures the impact on the fit of the model by considering all the individuals with the fit when deleting a particular individual from the sample.

Consider $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{(i)}$ the estimated maximum likelihood vectors from the sample with all individuals and the sample without individual i, respectively. An indicator of the influence of i-th individual can be calculated by $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}$. If the estimates differ substantially, the individual can be considered influential. A measure that also can be used and verifies the distance between the two likelihoods is the likelihood displacement, being given by

$$LD_i = 2\left[l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_{(i)})\right],$$

where $l(\hat{\boldsymbol{\theta}})$ and $l(\hat{\boldsymbol{\theta}}_{(i)})$ are, respectively, the likelihood logarithms of the parameters obtained from the sample with all points and the sample without the i-th individual. When it is not possible to obtain an analytical form for LD_i, it is necessary for an approximation method. Details about the measure of influence are covered in Cook and Weisberg (1982), McCullagh and Nelder (1989), Turkman and Silva (2000), Paula (2013), among others. It is highlighted that a point should only be excluded as a last alternative after several attempts to accommodate it in the fit, such as through transformations or including covariates (Silva, 2003).

2.5 Material and Methods

2.5.1 Material

As an application, it is considered the data from the experiment conducted by Marques (2018) regarding the histopathological alterations found in the livers of Tambaqui fish (*Colossoma macropomum*) at the Biofish-Aquicultura farm based in Porto Velho-RO from January 2015 to October 2016. In this experimental study, juvenile fish were anesthetized and marked using a microchip in the ventral portion (Figure 2.1), applying Methylene Blue to the inserted site to prevent infection. After recovering from anesthesia, the Tambaquis were managed in an excavated pond with approximately $600m^2$ of water, where they received the same food three times a day. In the end, the fish were fasted for 24 hours, collected with a trawl, and anesthetized when transported to water tanks for slaughter.

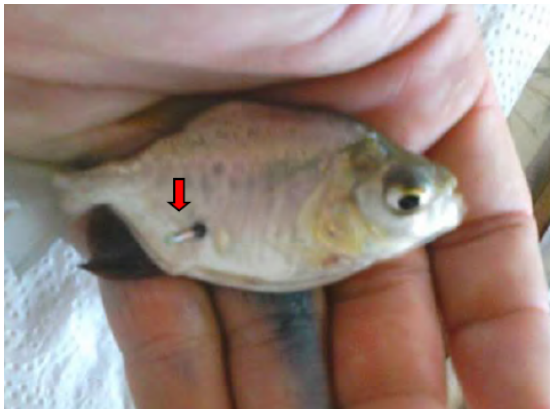

Figure 2.1. Microchip inserted in the juvenile of Tambaqui in a study carried out by Marques (2018) at the Biofish-Aquicultura farm.

The pituitary gland was collected for gene expression analysis, placed in a stabilizing solution (RNAlater), and stored at -80ºC until the moment of RNA extraction. With the DNA Analyzer 4300 equipment, two different types of genotypes, 122 and 130, were obtained. Small organ fragments were collected and properly stored for the liver histopathology analysis at the Laboratory of Ecology of Reproduction and Recruitment of Marine Organisms, Oceanographic Institut, USP/SP. The histopathological alterations were photomicrographed, Figure 2.2, and ordered according to the severity of the lesions, being classified as mild, moderate, and irreversible. Images of the lesions were obtained using the AXIOSKOP-ZEIS photomicroscope.

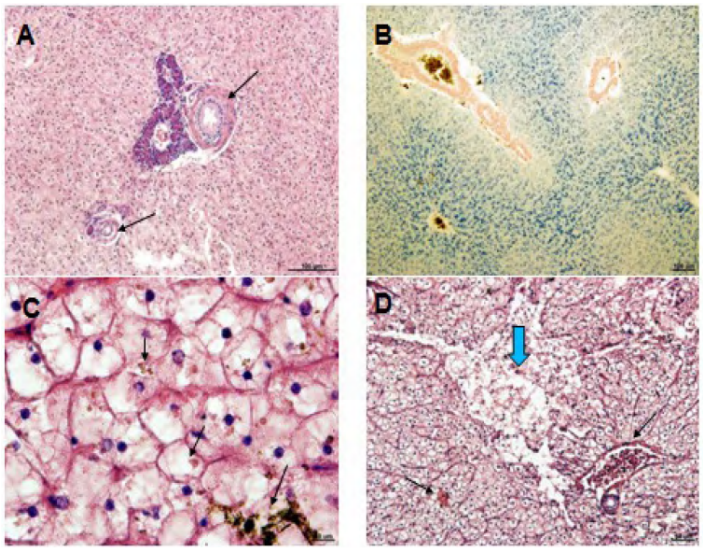

Figure 2.2. Morphology of the liver tissue of the Tambaqui fish with the histopathological alterations in the experiment carried out by Marques (2018) at the Biofish-Aquicultura farm. A- Ductal hypertrophy (black arrows); B- Hemosiderosis; C- Cholestasis (black arrows); D- Focal necrosis (blue arrow) and Congestion of vessels and sinusoids (black arrows).

The author made available 21 data from fish with genotype 122 and 21 from fish with genotype 130, totaling a sample of size equal to 42, in which was verified the relationship between the severity of lesions with the different gene expressions of Tambaqui. According to Marques (2018), the liver needs to function properly for a healthy fish population.

2.5.2 Methods

In this application, the response represents the histopathological alteration obtained in the liver of the fish associated with a different genotype, and the degree of severity of the lesions (from less to more severe) depends on this classification. Then, the variable response Y_i, $i = 1, 2, \ldots, 42$, has an ordinal scale assuming values in the set $\{1, 2, 3\}$, i.e., $Y_i = j$ represents the response of the i-individual in the category j, $j = 1, 2, 3$, where 1-mild, 2-moderate, 3-irreversible with $1 < 2 < 3$. In this context, the corresponding observed vector is $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3})'$, where $y_{ij} = 1$ if the response referring to the fish i belongs to the category j and $y_{ij} = 0$, otherwise. The genotype covariate is a factor, being incorporated into the model through the dummy variable.

First, to test the proportionality, the likelihood ratio test described in section 2.2.2 will be used considering the model with the main effect given by

$$\text{logit}\left[\gamma_{ij}(x_i)\right] = \log\left[\frac{\gamma_{ij}(x_i)}{1-\gamma_{ij}(x_i)}\right] = \alpha_j + \beta_j x_i, \quad j = 1, 2 \tag{2.6}$$

where α_j is the intercept, β_j is the parameter associated with the genotype effect on the j-th logit. Here, the third category is used as a reference. Using the standard parameterization, $x_i = 0$ for the i-th fish with genotype 122 and $x_i = 1$ for fish i with genotype 130.

If the proportionality condition is not violated, proportional odds are assumed. Otherwise, the model (2.6) is used to proceed with selecting the linear predictor. Under the proportionality assumption, the sequential proportional odds models are expressed by

Model 1 - Null:

$$\text{logit}\left[\gamma_{ij}(x_i)\right] = \log\left[\frac{\gamma_{ij}(x_i)}{1-\gamma_{ij}(x_i)}\right] = \alpha_j, \; j = 1, 2$$

Model 2 - Genotype effect:

$$\text{logit}\left[\gamma_{ij}(x_i)\right] = \log\left[\frac{\gamma_{ij}(x_i)}{1-\gamma_{ij}(x_i)}\right] = \alpha_j + \beta x_i, \quad j = 1, 2.$$

The likelihood ratio test (LRT) is used to select the structure of the linear predictor, verifying if there is an effect of genotype in the classification of severity found in the Tambaqui liver, that is, if $H_0 : \beta = 0$ is true or false. The test statistic is given by

$$\Lambda = -2\left[l_{H_0}(\hat{\boldsymbol{\alpha}}) - l_{H_1}(\hat{\boldsymbol{\alpha}}, \hat{\beta}),\right]$$

where $l_{H_0}(\hat{\boldsymbol{\alpha}})$ is the logarithm of the null model likelihood function and $l_{H_1}(\hat{\boldsymbol{\alpha}}, \hat{\beta})$ is the logarithm of likelihood function of the model with genotype effect, with expressions given by

$$l_{H_0}(\hat{\boldsymbol{\alpha}}) = \sum_{i=1}^{42}\sum_{j=1}^{3} y_{ij} \log\left(\frac{\exp(\hat{\alpha}_j)}{1+\exp(\hat{\alpha}_j)} - \frac{\exp(\hat{\alpha}_{j-1})}{1+\exp(\hat{\alpha}_{j-1})}\right)$$

and

$$l_{H_1}(\hat{\boldsymbol{\alpha}}, \hat{\beta}) = \sum_{i=1}^{42}\sum_{j=1}^{3} y_{ij} \log\left(\frac{\exp(\hat{\alpha}_j + \hat{\beta}x_i)}{1+\exp(\hat{\alpha}_j + \hat{\beta}x_i)} - \frac{\exp(\hat{\alpha}_{j-1} + \hat{\beta}x_i)}{1+\exp(\hat{\alpha}_{j-1} + \hat{\beta}x_i)}\right),$$

with $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \hat{\alpha}_2)'$. The estimates of the parameters of models (1) and (2) are obtained by the maximum likelihood procedure as described in the review chapter, section 2.2.2. The null model has only the intercept effect (2 parameters), and model 2 takes into account the intercept and genotype effect (3 parameters) under the null hypothesis has $\Lambda \sim \chi^2_1$.

Once the genotype effect is significant, confidence intervals (CIs) are constructed for the estimated probabilities for each response category and comparisons between observed and estimated proportions. In this way, simultaneous confidence intervals of $100(1-\alpha)\%$ are given by (see May and Johnson, 1997)

$$\hat{\pi}_{ij}(x_i) \pm \sqrt{\chi^2_{(\alpha,l)} \times \hat{\pi}_{ij}(x_i) \times [1 - \hat{\pi}_{ij}(x_i)]}, \quad j = 1, 2, 3$$

where $\chi^2_{(\alpha,l)}$ is the point from a chi-square distribution with $l = J-1 = 2$ degrees of freedom and $\alpha = 0,05$ is the significance level. The estimated probabilities are expressed by

$$\hat{\pi}_{i1}(x_i) = \frac{\exp(\hat{\alpha}_1 + \hat{\beta}x_i)}{1 + \exp(\hat{\alpha}_1 + \hat{\beta}x_i)},$$

$$\hat{\pi}_{i2}(x_i) = \frac{\exp(\hat{\alpha}_2 + \hat{\beta}x_i)}{1 + \exp(\hat{\alpha}_2 + \hat{\beta}x_i)} - \frac{\exp(\hat{\alpha}_1 + \hat{\beta}x_i)}{1 + \exp(\hat{\alpha}_1 + \hat{\beta}x_i)},$$

and

$$\hat{\pi}_{i3}(x_i) = 1 - \hat{\pi}_{i1}(x_i) - \hat{\pi}_{i2}(x_i).$$

Next step, for the fitted model validation, the surrogate residuals are used as described in section 2.3.4. Thus, with the data and the model, the conditional distribution of $Z_i \in (\hat{\alpha}_{j-1}; \hat{\alpha}_j)$ given $Y_i = j$ is obtained by substituting the parameter estimates $\hat{\alpha}_j$'s and $\hat{\beta}$ where the latent variable is $Z_i = -\hat{\beta}x_i + \varepsilon_i$ and $\varepsilon_i \sim Log(0, 1)$. A random sample s_i, $i = 1, 2, \ldots, 42$, is obtained from this distribution, and the i-th surrogate residual is given by

$$\hat{r}_i = s_i + \hat{\beta}x_i - \int_{-\infty}^{+\infty} udG(u).$$

Once obtained the residuals, it is possible to compare their empirical distribution function graphically with the standard logistic distribution function. Also, the bootstrap algorithm described in section 2.3.4 is used with 10 replications because of the sample size. The informal and formal techniques to evaluate the residual performance are the following: a) histogram, b) half-normal plot, c) the plot of residuals versus covariates, and c) the Kolmogorov-Smirnov test as described in section 2.4.

The analysis and estimation of model parameters were performed by the `clm(.)` function of the ordinal package (Christensen, 2013) and the `resids(.)` function of the sure package (Greenwell et al., 2018) to obtain the surrogate residuals. The ks.test(.) function of the `dgof` package (Arnold and Emerson, 2011) was used to obtain the p-value of the Kolmogorov Smirnov test. Finally, the `hnp(.)` function is used for the half-normal plot with a simulated envelope, implemented in the `hnp package` (Moral et al., 2017). All are available in the `R` software (R Core Team, 2020).

2.6 Results

Initially, an exploratory analysis was carried out to describe the fish data set. The frequencies of mild, moderate, and irreversible lesions were obtained for each type of genotype (Figure 2.3), in which one can observe the differences according to classifications. The liver alteration classified as irreversible had a higher frequency in fish with genotype 122 than in fish with genotype 130. On the other hand, fish with genotype 130 had higher frequencies of mild and moderate lesions than fish with genotype 122.

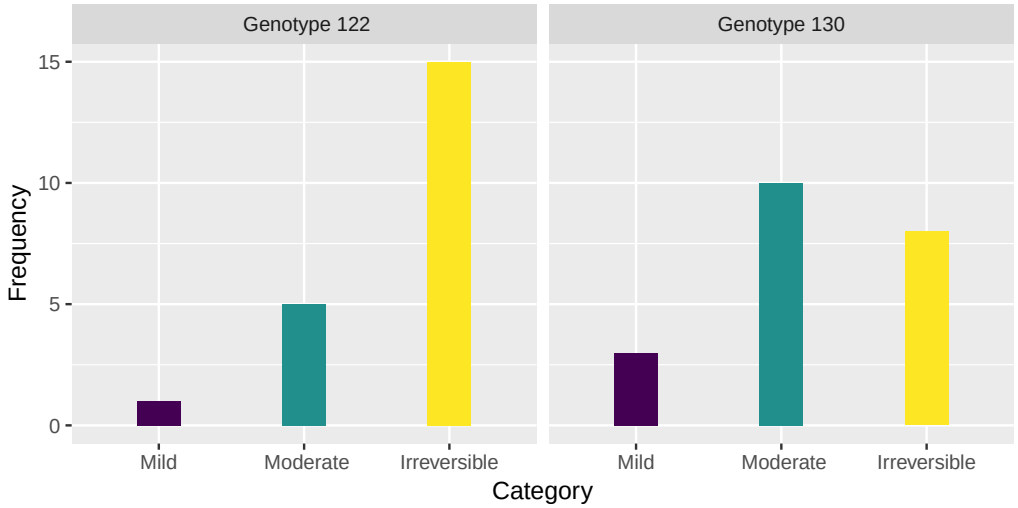

Figure 2.3. Frequencies of mild, moderate, and irreversible lesions in the liver of Tambaquis by type of genotype (122 and 130) in the study carried out by Marques (2018) at the Biofish-Aquicultura farm.

Then the cumulative logit and proportional odds models were fitted to test proportionality. It was verified evidence in favor of the proportional odds model by the LRT (p-value = 0.8667). Afterward, the sequential proportional odds models were fitted and compared using the LRT as well. The model that considers the genotype effect was selected (p-value= 0.02714). Based on this result, it is concluded that the type of genotype contributes to explaining the lesion classification in the liver of the Tambaqui fish in the study carried out by Marques (2018).

The estimated parameters and standard errors for the model with genotype effect are presented in Table 2.1.

Table 2.1. Estimated regression parameters of the proportional odds model with the effect of genotype selected for analysis Tambaqui in a study carried out by Marques (2018).

Parameter	Estimate	Standard error
α_1 (intercept 1)	-3.1289	0.6989
α_2 (intercept 2)	-0.9079	0.4811
β (Genotype 133)	1.3779	0.6437

The expressions in terms of the cumulative logits for the proportional odds model with genotype effect are expressed by

$$\log\left[\frac{\gamma_1(x)}{1-\gamma_1(x)}\right] = -3.1289 + 1.3779\,x \quad \text{and} \quad \log\left[\frac{\gamma_2(x)}{1-\gamma_2(x)}\right] = -0.9079 + 1.3779\,x.$$

The interpretation of the estimated parameter is generally performed through the odds ratios. The estimate of the genotype effect parameter is 1.3779 (Table 2.1), which indicates a tendency towards classification in the less severe categories in fish with genotype 130, as observed in the exploratory analysis. Therefore, the odds of the lesion being classified as mild (in relation to moderate or irreversible) in fish with genotype 130 was approximately 3.97 times the odds of being classified in fish with genotype 122. The same conclusions can be obtained considering the odds of the lesion being classified as mild or moderate in relation to irreversible, which occurs due to the proportionality assumption assumed by the model.

The predicted probabilities for each response category in the different types of genotype, with their respective confidence intervals, are presented in Table 2.2. Fish with genotype 122 showed irre-

versible liver alteration with a probability of 71.26%, while for fish with genotype 130, this occurs with a probability of 38.46%. Therefore, fish with genotype 122 tend to have more severe liver lesions than fish with genotype 130. As shown in Table 2.2, the confidence interval has greater amplitude due to the relatively small sample size.

Table 2.2. Estimated probabilities and 95% confidence intervals (in parentheses) in each of the response categories for fish with genotypes 122 and 130 were obtained by fitting the proportional odds model with the genotype effect.

Genotype	Category		
	mild	moderate	irreversible
122	4.19%	24.55%	71.26%
	(1.10%; 14.69%)	(11.87%; 44.02%)	(49.13%; 86.42%)
130	14.79%	46.75%	38.46%
	(5.41%; 34.49%)	(29.22%; 65.12%)	(20.94%; 59.59%)

The observed and estimated proportions by genotype can be seen in Figure 2.4. Visually, the values are close to each other, showing that the proportional odds model includes the genotype effect is reasonable for describing the proportions of lesions observed in the study conducted by Marques (2018).

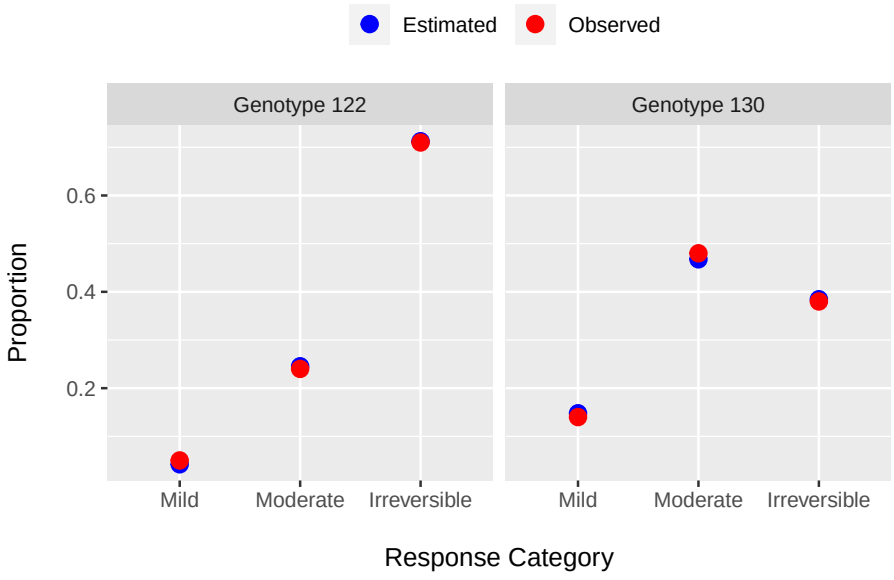

Figure 2.4. Observed proportions for the mild, moderate, and irreversible lesions and proportions estimated by proportional odds model with genotype effect in the study of Marques (2018).

The validation of the model assumptions was verified by the surrogate residuals analysis using bootstrap replications due to the sample size. Observing the histogram, Figure 2.5, the residual distribution presented a shape similar to the standard logistic distribution, which is symmetrical, similar to the normal distribution but with heavier tails. The values for mean and variance were approximately 0.002 and 3.176, respectively. Furthermore, the p-value of the Kolmogorov-Smirnov test was approximately 0.729, which indicates in favor of the hypothesis that the surrogate residuals follow a standard logistic distribution.

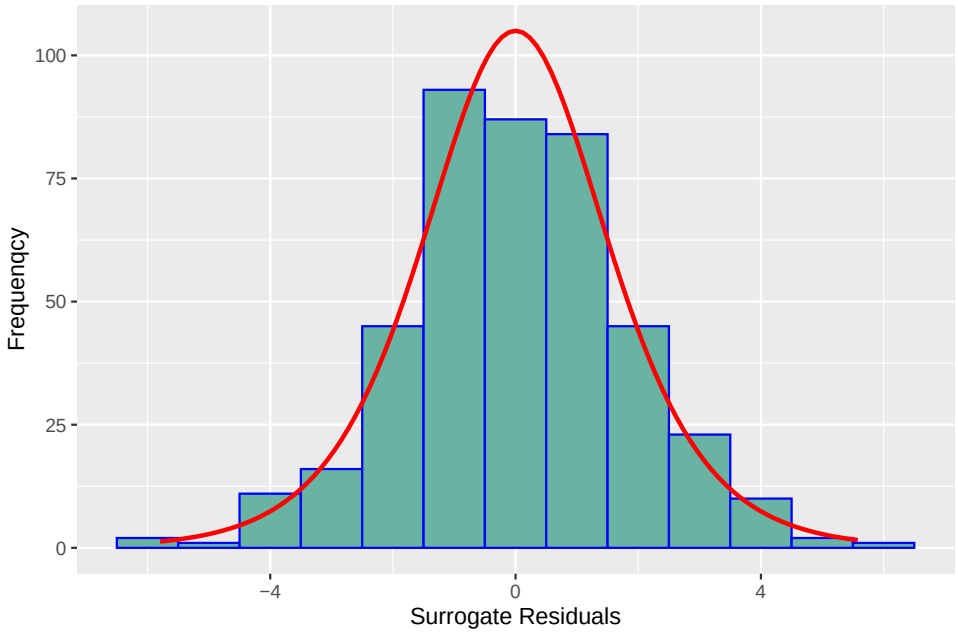

Figure 2.5. Histogram of surrogate residuals related to the proportional odds fitted model (genotype effect) to the fish data in the study of Marques (2018)

The half-normal plot with a simulated envelope for the surrogate residuals was presented in Figure 2.6. There is evidence that the observed data are a plausible realization of the fitted model since no systematic deviation pattern was observed with all the points inside the envelope. Thus, the model with the genotype effect can be used to analyze the data.

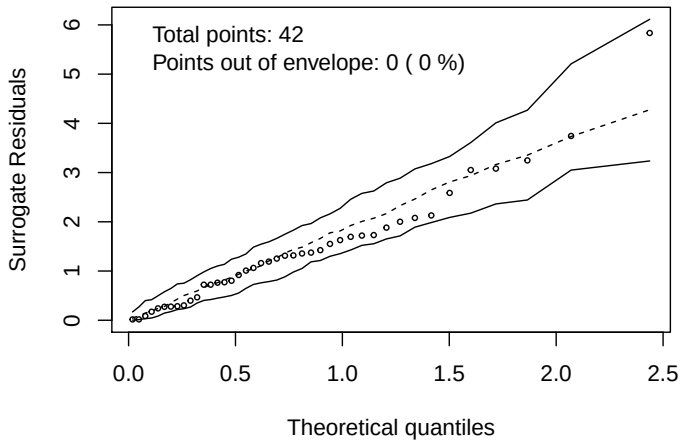

Figure 2.6. Half-normal plot with a simulated envelope for the surrogate residuals to assess the fit of the model with genotype effect in the study of Marques (2018).

As in this model, a covariate is a factor, using the plot of residuals versus covariate is inappropriate. The boxplot of residuals was obtained for each genotype (Figure 2.7), which revealed medians of residuals close to zero. In addition, the residual distributions present symmetrical tendency, similar variability, and the presence of outliers per genotype.

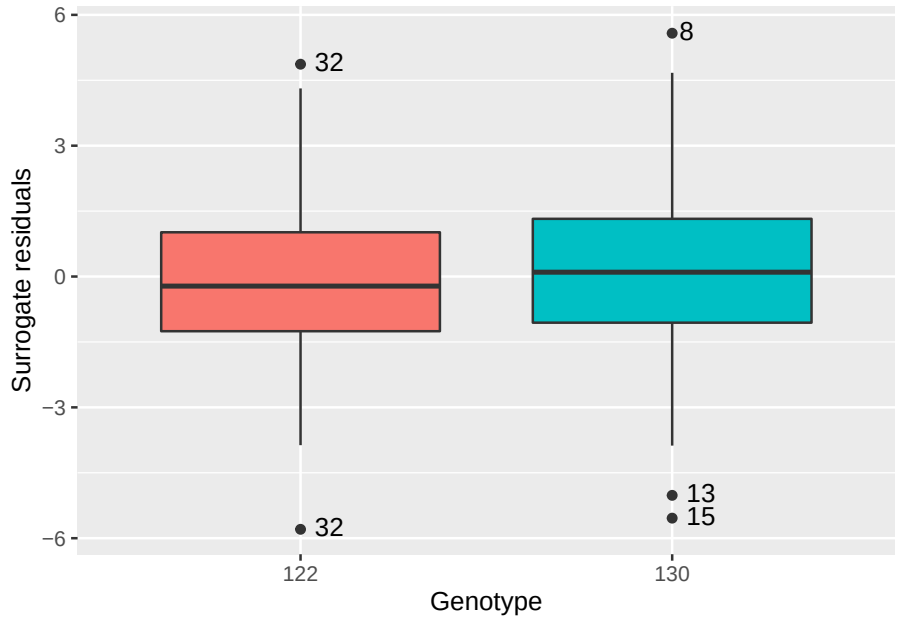

Figure 2.7. Boxplot of surrogate residuals per genotype to assess the proportional odds fitted model to the fish data in the study of Marques (2018).

The large residuals in the Figure 2.7 refer to individual #32 for genotype 122 and to individuals #8, #13, and #15 for genotype 130. The model was fitted without these individuals to assess the impact on the estimates of model parameters. The parameter estimates and the related standard errors, in parentheses, are shown in Table 2.3.

Table 2.3. Estimated Parameters of proportional odds model with genotype effect by excluding the individual #32 for genotype 122 and the individuals #8, #13 and #15 for genotype 130 the fish data.

Individual	Parameters		
	α_1	α_2	β
Complete sample	-3,1289	-0,9079	1,3779
	$(0, 6989)$	$(0, 4811)$	$(0, 6437)$
Excluding #32	-3,0651	-0,8391	1.3105
	$(0, 7003)$	$(0, 4857)$	$(0, 6467)$
Excluding #8	-3,0358	-0,9137	1,3142
	$(0, 6928)$	$(0, 4806)$	$(0, 6485)$
Excluding #13	-3,3315	-0,8971	1,2675
	$(0, 7532)$	$(0, 4822)$	$(0, 6501)$
Excluding #15	-3,3315	-0,8971	1,2675
	$(0, 7532)$	$(0, 4822)$	$(0, 6501)$

The variations between the estimated parameters (and the standard errors) were not disproportionate with the exclusion of individuals by genotype from the sample (Table 2.3), indicating that these points do not have a high influence on the fit. Thus, the entire inference based on the complete sample remains valid, and the choice of another model could lead to inadequate conclusions. Finally, the results were satisfactory, contributing to the validation of the model that provided a good fit for the data.

2.7 Conclusion

The paper describes an introduction to residuals analysis with ordinal data through a method that uses a continuous variable that replaces the original response, allowing to obtain unique residuals by

individuals. The surrogate residuals have similar properties to ordinary residuals for a continuous response and they can be used in virtually all available diagnostic tools, as illustrated in the practical application. The residuals were informative, not detecting violations of the assumptions of the model selected to describe the fish data. As the residuals are obtained by conditional sampling, it is recommended to use the Bootstrap algorithm in small samples to control the sampling error that can lead to a variation in the patterns of residuals. The limitation of this approach is that the residual is defined only for models that present a valid proportional odds assumption, not covering the entire class of models for ordinal data. Furthermore, these univariate residuals are not defined for nominal data or the different data structure from the individual one. These issues present challenges in the diagnostics for different models with distinct data structures. Future studies can be carried out to improve the analysis of residuals in polytomous data, stimulating the methodological development in this important area whose tools are still limited.

References

Agresti, A. (2002). *An introduction to categorical data analysis*. John Wiley & Sons, Nova Jersey, 3 edition.

Agresti, A. (2007). *An introduction to categorical data analysis*. John Wiley & Sons, Hoboken, New Jersey, 2 edition.

Agresti, A. (2010b). *Analysis of ordinal categorical data*. John Wiley & Sons, Nova Jersey, 2 edition.

Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, 26(6):1323–1333.

Arbogast, P. G. and Lin, D. (2005). Model-checking techniques for stratified case-control studies. *Statistics in medicine*, 24(2):229–247.

Arnold, T. B. and Emerson, J. W. (2011). Nonparametric goodness-of-fit tests for discrete null distributions. *R Journal*, 3(2).

Atkinson, A. C. (1985). Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis. Technical report.

Bilder, C. R. and Loughin, T. M. (2014). *Analysis of categorical data with R*. Chapman and Hall/CRC Press, Boca Raton, 1 edition.

Christensen, R. H. B. (2013). ordinal: Regression models for ordinal data. *R package version*, 28:56.

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

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.

Correa, R. O., Souza, A. R. B., and Martins Junior, H. (2018). Criação de tambaquis. *Embrapa Amazônia Oriental-Fôlder/Folheto/Cartilha (INFOTECA-E)*.

Dufour, J. M., Farhat, A., Gardiol, L., and Khalaf, L. (1998). Simulation-based finite sample normality tests in linear regressions. *The Econometrics Journal*, 1(1):154–173.

Efron, B. (1979). Bootstrap method: another look at the jackknife. *The annals of statistics*, 7:1–26.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.

Giolo, S. R. (2017). *Introdução à análise de dados categóricos com aplicações*. Editora Blucher, São Paulo, 1 edition.

Greenwell, B. M., McCarthy, A., Boehmke, B. C., and Liu, D. (2018). Residuals and diagnostics for binary and ordinal regression models: An introduction to the sure package. *The R Journal*, 10(1):381–394.

Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Giornali dell'Istituto Italiano degli Attuari*, 4:83–91.

Lemos, T. D. O., Rodrigues, M. D. C. P., De Lara, I. A. R., De Araújo, A. M. S., De Lemos, T. L. G., Pereira, A. L. F., and De Paula, L. V. T. (2015). Modeling the acceptability of cashew apple nectar brands using the proportional odds model. *Journal of Sensory Studies*, 30(2):136–144.

Li, C. and Shepherd, B. E. (2012). A new residual for ordinal outcomes. *Biometrika*, 99(2):473–480.

Liu, D. and Zhang, H. (2018). Residuals and diagnostics for ordinal regression models: A surrogate approach. *Journal of the American Statistical Association*, 113(522):845–854.

Liu, I., Mukherjee, B., Suesse, T., Sparrow, D., and Park, S. K. (2009). Graphical diagnostics to check model misspecification for the proportional odds regression model. *Statistics in medicine*, 28(3):412–429.

Lopes, I. G., De Oliveira, R. G., and Ramos, F. M. (2016). Perfil do consumo de peixes pela população brasileira. *Biota Amazônia (Biote Amazonie, Biota Amazonia, Amazonian Biota)*, 6(2):62–65.

Marques, M. F. (2018). *Associaçãao de polimorfismo microssatélite no gene GH em Tambaqui (Colossoma macropomum) com caracterísstica fenotípicas e expressão gênica*. PhD thesis, Universidade de São Paulo.

May, W. L. and Johnson, W. D. (1997). Properties of simultaneous confidence intervals for multinomial proportions. *Communications in Statistics-Simulation and Computation*, 26(2):495–518.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2 edition.

Moral, R. A., Hinde, J., and Demétrio, C. G. B. (2017). Half-normal plots and overdispersed models in r: the hnp package. *Journal of Statistical Software*, 81(1):1–23.

Ng, K. W., Tian, G. L., and Tang, M. L. (2011). *Dirichlet and related distributions: Theory, methods and applications*.

Paula, G. A. (2013). *Modelos de regressão: com apoio computacional*. IME-USP São Paulo.

Peterson, B. and Harrell Jr, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39(2):205–217.

R Core Team, . (2020). R: A language and environment for statistical computing.

Reiter, J. P. and Kohnen, C. N. (2005). Categorical data regression diagnostics for remote access servers. *Journal of Statistical Computation and Simulation*, 75(11):889–903.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Silva, J. A. P. (2003). *Métodos de diagnóstico em modelos logísticos trinomiais*. Dissertação (mestrado em estatística).

Singer, J. M., Rocha, F. M. M., and Nobre, J. S. (2017). Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review*, 85(2):290–324.

Souza, E. C. (2006). Análise de influência local no modelo de regressão logística.

Turkman, M. A. A. and Silva, G. L. (2000). Modelos lineares generalizados–da teoria à prática. *Sociedade Portuguesa de Estatística, Lisboa*, page 153.

Tutz, G. (2011). *Regression for categorical data*, volume 34. Cambridge University Press, Cambridge.

Williams, O. D. and Grizzle, J. E. (1972). Analysis of contingency tables having ordered response categories. *Journal of the American Statistical Association*, 67(337):55–63.

Appendix

```
#########################################################################
# R code #
#########################################################################
#Fish data
#ordinal variable
rm(list=ls(all=TRUE))
# Installing the packages
library(ordinal);library(hnp);library(ggplot2); library(sure);
library(gridExtra);library(dgof)
#########################################################################################
mydata<-read.csv("fish.csv", head=TRUE, sep=";", dec=",") #reading data
mydata$genotype<-as.factor(mydata$genotype) #covariate
mydata$resp<-as.ordered(as.factor(mydata$resp)) #ordinal response
attach(mydata)
summary(mydata)
head(mydata)
#########################################################################################
#Exploratory analysis
levels(mydata$genotype)<-c(" Genotype 122","Genotype 130")
levels(mydata$resp)<-c("Mild", " Moderate", "Irreversible")
ggplot(mydata, aes(x = resp, fill = resp)) +
geom_bar(width=0.3,show.legend = FALSE) + facet_grid(.~genotype)+
ylab('Frequency')+xlab("Category")
#########################################################################################
#Models
mod <- clm(resp~genotype,data=mydata) # MOP
#Likehood ratio test
nominal_test(mod)
#or
mod1 <- clm(resp~genotype,nominal=~genotype,data=mydata) # MLC
anova(mod,mod1)
#########################################################################################
#Likehood ratio test to select linear predictor of sequential proportional
odds models
mod0 <- clm(resp~1,data=mydata)
anova(mod0,mod)
#Deviances
tab <- with(mydata, table(genotype, resp))
pi.hat <- tab/rowSums(tab)
(logvero_modc <- sum(tab * ifelse(pi.hat > 0, log(pi.hat), 0)))
logvero_mod0 <- mod0$logLik
(Deviance0 <- -2 * (logvero_mod0 - logvero_modc))
logvero_mod <- mod$logLik
(Deviance1 <- -2 * (logvero_mod - logvero_modc))
#########################################################################################
```

```
#Wald CI 95% for:
#parameters
param<-coefficients(mod) #coefficients of parameters
confint(mod,type = "Wald")
# and the estimated probabilities
drop<-expand.grid(genotype=levels(mydata$genotype))
CIprob<-predict(mod,newdat=drop,se.fit=TRUE,interval = T)
#odds ratio
exp(-param[3])
##############################################################################
#observed versus estimated probalities plot
tab <- with(mydata, table(genotype, resp)) #frequency
prob<-round(prop.table(tab,margin = 1),2);prob #observed probabilities
p1<-as.vector(t(prob))
probs<-as.vector(t(predict(mod, newdat=drop)$fit)) #estimated probabilities
probfinal<-data.frame(genotype=rep((1:2),each=3,times=2),
                      response=rep((1:3),times=4))
datafinal<-cbind(probfinal,proba=c(probs,p1),tipo=rep(c("Estimated",
"Observed"),each=6))
datafinal$genotype<-as.factor(datafinal$genotype)
datafinal$response<-as.factor(datafinal$response)
levels(datafinal$response)<-c("Mild","Moderate","Irreversible")
levels(datafinal$genotype)<-c("Genotype 122", "Genotype 130")
ggplot(datafinal, aes(x=response, y=proba, colour=tipo)) +
  geom_point(size=3) + facet_grid(.~genotype) +
  xlim("Mild","Moderate","Irreversible")+
  xlab('\n Response Category \n')+ ylab('Proportion\n')+
  scale_colour_manual(name="",breaks=c('Estimated','Observed'),
                      values=c('blue','red'))+ theme(legend.position="top")
##############################################################################
#hnp using surrogate residuals
res_sure<-resids(mod,nsim = 10) #to obtain the residuals
#half-normal plot with simulated envelope
hnp(res_sure,print=T, ylab="Surrogate Residuals",scale = T)
#QQ plot
qq_sure <- autoplot.clm(mod, nsim = 10, what = "qq");qq_sure
#The function to obtain the bootstrap sample
nsim<-10 # number of replications
n.obs<-mod$nobs #sample size
boot.res <- boot.index <- matrix(nrow = n.obs, ncol = nsim)
for(i in seq_len(nsim)) {
  boot.index[, i] <- sample(n.obs, replace = TRUE)
  mr<- mod$y[boot.index[, i]]
  boot.res[, i] <- resids(mod, y = y[boot.index[, i]], mean.response = mr)
}
x_orig<-as.vector(boot.index)
xboots<-vector()
```

```
for(i in 1:length(x_orig)) {
 if(x_orig[i]<=21){
   xboots[i]<-"130"
 }else{
   xboots[i]<-"122"
 }
}
yboots<-as.vector(boot.res)
mydataboots<-data.frame(xboots,yboots)
attach(mydataboots)
#p-value of Kolmogorov-Smirnov Test for bootstrap residuals
ks.test(yboots, "plogis")$p.value
#mean and standad deviation of bootstrap residuals
mean(yboots); sd(yboots)^2
#Boxplot dos resíduos com 10 rep bootstrap
(p10 <- ggplot(mydataboots, aes(x =xboots,y = yboots))+labs(x = "Genotype",
y = "Surrogate residuals")+ geom_boxplot(aes(fill=xboots))+
guides(fill=FALSE))
#to obtain the outliers per genotype
out <- ggplot_build(p10)[["data"]][[1]][["outliers"]]
g122.out<-as.vector(out[[1]])
g130.out<-as.vector(out[[2]])
ind_boots<-match(c(g122.out,g130.out), yboots)
ind_orig<-x_orig[ind_boots]
out_f<-rep(NA,length(x_orig))
for(i in 1:length(x_orig)){
  for (j in 1:length(ind_orig)) {
    if(i==ind_boots[j]){ out_f[i]<-ind_orig[j]}}}
#Boxplot with the individuals that corresponds the outliers per genotype
(p10+ geom_text(aes(label=out_f),na.rm=TRUE,nudge_y=0.05,hjust=-0.5))

#Histogram with 10 replicates bootstrap
ggplot(mydataboots, aes(x=yboots)) + geom_histogram(binwidth=1,
fill="#69b3a2", color="blue")+ylab("Frequenqcy")
+xlab("Surrogate Residuals") +stat_function(fun = function(x)
dlogis(x, 0,1)*length(yboots),color = "red", size = 1)
##############################################################################
```

3 NOMINAL DATA AND DIAGNOSTICS BASED ON RANDOMIZED QUANTILE RESIDUALS AND DISTANCE MEASURES

Abstract

Nominal variables are of interest in research in many areas of knowledge. Depending on the study objective, these data can be obtained from experiments with an individual or grouped structure. The generalized logit model is commonly used to relate the potential effects of covariates on response. After fitting a multi-categorical model, one of the challenges is the definition of an appropriate residual and choosing diagnostic techniques, which are still under development in the scientific area. As the response variable is multivariate, the ordinary residual is a vector for each individual with asymptotic distribution generally unknown. The definition of an appropriate residual enables the correct analysis in diagnostic tools. In this context, this work assesses the normality of the randomized quantile residual associated with the individual nominal data and proposes to identify the presence of outliers through Euclidean and Mahalanobis distances by reducing the ordinary residual dimension associated with the grouped data. These methodologies were used in diagnostic techniques for assessing the generalized logit models through simulation studies, whose results attest to the good performance of their application. Two data sets of literature were presented to illustrate these methods. The parameters estimation was performed via maximum likelihood, and the residuals and the values of distances were analyzed via a half-normal plot and Shapiro-Wilk test. Overall, it was possible to check the model assumptions, which provided evidence that the observed data were plausible realizations of the fitted models.

Keywords: Generalized logit model; Maximum likelihood; Half-normal plot; Shapiro-Wilk test.

3.1 Introdution

Nominal polytomous variables are defined by a finite set of categories (more than two), being of interest in experiments in several scientific areas such as agricultural, biological, and others. For example, in agricultural sciences, experiments are designed in which the experimental unit is an individual (a plant, an insect, or an animal), recording the categorized response. However, practical situations are not rare in which the experimental unit is composed of a fixed group of individuals, such as a stall with animals, a cage with insects, or a plant with its branches, among others, for these cases are considered categorized data terms a grouped structure.

The generalized logit model is frequently used for the statistical analysis of nominal data with individual or grouped structures, simultaneously describing the relationship between the probabilities for all pairs of response categories with the covariates of the study (Agresti, 2002). On the other hand, the assumptions of the fitted model must be verified to validate the statistical inference, being the residual analysis fundamental in this process. So, the first step is the definition of an appropriate residual that can be used in formal (tests) and informal (graphs) diagnostic techniques to assess the goodness-of-fit and model assumptions. According to Feng et al. (2020), residuals are essential in identifying discrepancies between the model and the data, detecting outliers and influential points. However, the analysis of residuals is a challenge for the multinomial case. As the response variable is multivariate, the ordinary residual defined by the difference between the observed response and the estimated probability is a vector for each individual, with a dimension defined by the number of categories. This residual has an asymptotic distribution unknown, making it difficult to interpret in diagnostic graphs (Reiter and Kohnen, 2005). In addition, deviance and Pearson statistics are quantitative measures widely used to test the goodness-of-fit of generalized linear models (GLMs). Still, they can only be applied to multinomial data in the grouped structure and with restrictions on the sample size (Tutz, 2011).

In this context, it is important to find or adapt techniques to overcome these limitations and could be considered some alternatives for diagnostics in these cases. The first is to reduce the number of categories (grouping into two) and perform residual analysis for the logistic regression model, whose techniques are consolidated in the statistical literature (Pregibon, 1981, Landwehr et al., 1984, Hossain and Islam, 2003, among others). However, grouping categories leads to the loss of information, changing the original questions of scientific research. Another alternative would be to fit the generalized logit model separately and define residuals for each sub-model, applying them to the different diagnostic tools. Silva (2003) presented deviance residuals without sign assignment and the Pearson residuals for the generalized logit sub-models with three categories, examining their performance in the plots of residuals versus predicted probabilities and from residuals versus the order of observations. However, the maximum likelihood estimates from the separate fit differ from those obtained in simultaneous fit, and their standard errors tend to be larger (Agresti, 2002).

Continuing within this framework, Cheng et al. (2021) defined a continuous residual vector for the individual nominal case based on the methodology of Liu and Zhang (2018), in which a uni-dimensional residual was described to individual ordinal data. The residuals were evaluated in the different dimensions for diagnosing the generalized logit model with three categories. The authors also presented the deviance and Pearson residuals vectors for comparison in some scenarios. These exhibited nearly parallel curves in the residuals versus covariate plots, while the continuous residuals showed expected patterns for the correct model. However, if the number of categories increases, the residual dimension increases, and the number of values can make it difficult to interpret both residuals and diagnostics.

On the other hand, for nominal data with grouped structure, Seber and Nyangoma (2000) defined a vector of residuals with basis on so-called projected residuals presented by Cook and Tsai (1985) in the nonlinear regression. The authors examined their proximity to the normal distribution and the magnitude of bias term associated with each residual to assess the fit of the log-linear model in examples from genetics and psychology. Furthermore, the Pearson residual vector was presented by Gupta et al. (2008) to detect influential points in the fit of the generalized logit model, whose parameters were estimated using the minimum phi-divergence estimator. However, these methodologies require theoretical development and are not implemented in statistical software.

A residual defined for a broad class of models that can be easily implemented in statistical software is the randomized quantile residual (Dunn and Smyth, 1996), an alternative for diagnostics associated with generalized logit models. Even so, a lack of investigation of its performance through simulation studies is observed for polytomous regression models. This residual for discrete data is an extension of the quantile residual to continuous data, in which randomization between two consecutive distribution functions has been introduced to produce continuous residuals. Randomized quantile residuals follow an approximately normal distribution if the estimated parameters are consistent, but it is important to investigate their properties in small sample sizes (Pereira, 2019). Feng et al. (2020) investigated the randomized quantile residuals for count data by comparing them with Pearson and deviance residuals. Through simulation studies, the authors concluded that the randomized quantile residuals are better approximated by the standard normal distribution than the others, detecting a lack of fit in the Shapiro-Wilk test and diagnostic plots for models associated with the data.

Another alternative is to propose distance metrics, such as Euclidean and Mahalanobis, to reduce the dimension of the ordinary residual for the diagnostics of generalized logit models. These metrics are widespread in the literature on Multivariate Analysis, calculating how far two individuals are in the original variable space in different analyses, for example, principal components, cluster analysis, and others (Johnson and Wichern, 2007). In the context of diagnostics, it is observed that these distances have already been used to detect outliers in linear regression (Hadi et al., 2009 and Ghorbani, 2019). However, there are no records of their use in models for nominal data.

In this chapter, the objectives are to assess the normality of randomized quantile residual and propose using Euclidean and Mahalanobis distances to reduce the dimension of ordinary residual in the diagnosis of generalized logit models associated with nominal data with both structures, individual and grouped. More specifically, it conducts simulation studies to (1) demonstrate that the randomized quantile residuals approximately follow a standard normal distribution for the correctly specified generalized logit model to the individual nominal data, examining them by the Shapiro-Wilk test, and (2) examine the performance of the values resulting from reducing the dimension of the ordinary residual by distance measures under the half-normal plot to detect outliers in the diagnosis of the generalized logit model in grouped data analysis.

The sections are organized as follows. The review of models and residuals for nominal polytomous data are presented in Section 3.2. The definition of the randomized quantile residual and the distance metrics (Euclidean and Mahalanobis) are presented in Sections 3.3 and 3.4, respectively. Section 4.3 presents the framework based on randomized quantile residuals and distances for nominal responses, which are the contributions of this chapter. The simulation studies and results are presented in Section 4.5. Two applications from literature to illustrate the methodologies and their results are presented in Section 4.5. Finally, the conclusion is presented in section 3.8.

3.2 Review of models and residuals for nominal polytomous data

Statistical models are based on the probability distribution of the response variable, with the multinomial distribution being assumed for polytomous data (nominal or ordinal). The set of assumptions, response distribution, and covariates (linear predictor structure) are essential in defining the model, influencing the construction of residuals, and applying diagnostic techniques.

3.2.1 Multinomial Distribution

Let it be a multinomial trial, that is, an experiment that admits J possible and mutually exclusive outcomes, whose probabilities are denoted by $\pi_1, \pi_2, ..., \pi_J$ such that $0 \leq \pi_j \leq 1$, $j = 1, 2, ..., J$, and $\sum_{j=1}^{J} \pi_j = 1$.

Consider m identical and independent trials, which means that the probabilities of occurrence of the results are constant for each trial and that the result obtained in one trial does not interfere with the result of the other. Taking the random variable Y_j, which represents the number of times the index j was observed in m trials, then the random vector $\mathbf{Y} = (Y_1, \ldots, Y_J)'$ follows a multinomial distribution with parameters m and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)'$, $\mathbf{Y} \sim Multi(m, \boldsymbol{\pi})$, and probability mass function given by

$$f(\mathbf{y}; \boldsymbol{\pi}) = P(Y_1 = y_1, Y_2 = y_2, ..., Y_J = y_J; m; \boldsymbol{\pi}) = \frac{m!}{y_1! y_2! ... y_J!} \pi_1^{y_1} \pi_2^{y_2} ... \pi_J^{y_J},$$

where $y_j \in \{0, 1, \ldots, m\}$ and $\sum_{j=1}^{J} y_j = m$.

For the category j the result y_j has mean and variance given by $\mathrm{E}(Y_j) = m\pi_j$ e $\mathrm{Var}(Y_j) = m\pi_j(1 - \pi_j)$, respectively. Furthermore, the covariance between y_j and y_k, $\forall j \neq k$, $j, k = 1, \ldots, J$, is obtained by $\mathrm{Cov}(Y_j, Y_k) = -m\pi_j\pi_k$, and that the marginal distribution of each y_j is binomial.

The multinomial distribution belongs to the canonical multiparametric exponential family, with a vector of canonical parameters $\boldsymbol{\theta} = [\log(\pi_1), \ldots, \log(\pi_J)]'$ and canonical statistics $\mathbf{T} = (Y_1, \ldots, Y_J)'$. However, the minimum representation of the exponential family is obtained considering the vectors $\boldsymbol{\theta} = \left[\log\left(\frac{\pi_1}{\pi_J}\right), \ldots, \log\left(\frac{\pi_{J-1}}{\pi_J}\right)\right]'$ and $\mathbf{T} = (Y_1, \ldots, Y_{J-1})'$, both of dimension $J - 1$, due to the restriction $\sum_{j=1}^{J} \pi_j = 1$, which results in the multiparametric exponential family with dimension $J - 1$ expressed in

the following form

$$f(\mathbf{y};\boldsymbol{\theta}) = \frac{m!}{y_1!\dots y_J!}\exp\left[\sum_{j=1}^{J-1}\theta_j y_j - b(\boldsymbol{\theta})\right],$$

where $\theta_j = \log\left(\frac{\pi_j}{\pi_J}\right)$, $j = 1,\dots,J-1$, and $b(\boldsymbol{\theta}) = m\log\left(1+\sum_{j=1}^{J-1}e^{\theta_j}\right)$.

According to Agresti (2007), the multinomial distribution is the most used in the class of generalized models for polytomous responses.

3.2.2 Nominal data structures

To present the notation of individual data, consider the response of an individual i, $i = 1,2\dots,n$, in one of the J categories. Let the indicator random variable $Y_{ij} = 1$ if the response of individual i is in category j, $j = 1,2,\dots,J$, and $Y_{ij} = 0$ otherwise, with $\sum_{j=1}^{J} Y_{ij} = 1$. Then, the random vector referring to the individual i given by $\mathbf{Y}_i = (Y_{i1},\dots,Y_{iJ})'$ has a multinomial distribution, $\mathbf{Y}_i \sim Multi(1,\boldsymbol{\pi}_i)$, with probability vector $\boldsymbol{\pi}_i = \mathrm{E}(\mathbf{Y}_i) = (\pi_{i1},\dots,\pi_{iJ})'$ where $\sum_{j=1}^{J}\pi_{ij} = 1$.

In the case of grouped data, let be the i-th experimental unit, $i = 1,2\dots,n$, composed of a group of individuals with a fixed size equal to m_i. The random variable given by Y_{ij} represents the number of times category j was observed in m_i individuals, with $\sum_{j=1}^{J} Y_{ij} = m_i$. Then, the random vector $\mathbf{Y}_i = (Y_{i1},\dots,Y_{iJ})'$ follows a multinomial distribution, $\mathbf{Y}_i \sim Multi(m_i,\boldsymbol{\pi}_i)$, with parameters m_i and $\boldsymbol{\pi}_i = (\pi_{i1},\dots,\pi_{iJ})'$. These data are usually arranged in a contingency table that, generically, can be represented by Table 3.1, in which the counts observed in the cells are represented by y_{ij}, $i = 1,2,...,n$ and $j = 1,2,...,J$.

Table 3.1. Example of a generic contingency table with n experimental units and J response categories

Experimental unit (i)	Response categories (j)					Total
	1	2	3	...	J	
1	y_{11}	y_{12}	y_{13}	...	y_{1J}	m_1
2	y_{21}	y_{22}	y_{23}	...	n_{2J}	m_2
...	...	...	...	...	...	...
n	y_{n1}	y_{n2}	y_{n3}	...	y_{nJ}	m_n

3.2.3 Generalized logit model

Consider a random sample of dimension n where for each experimental unit i, $i = 1,2,\dots,n$, is associated with a vector of covariates over the nominal response. The model, which compares each category with one chosen as a reference, is defined as

$$\text{logit}\left[\pi_{ij}(\mathbf{x}_i)\right] = \log\left[\frac{\pi_{ij}(\mathbf{x}_i)}{\pi_{iJ}(\mathbf{x}_i)}\right] = \alpha_j + \sum_{k=1}^{p}\beta_{jk}x_{ik} = \alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i, \quad j = 1,\dots,J-1, \tag{3.1}$$

where J is the number of categories, $\pi_j(\mathbf{x}_i)$ is the probability of response of individual i in the j-th category, $\mathbf{x}_i = (x_{i1},x_{i2},\dots,x_{ip})'$ is the vector of p covariates, $\boldsymbol{\beta}_j = (\beta_{j1},\beta_{j2},\dots,\beta_{jp})'$ represents the parameter vector, and α_j is the intercept. According to Agresti (2002), the covariates can be quantitative, factors (using dummy variables) or both.

The equations that express the model directly in terms of the probabilities are

$$\pi_{ij}(\mathbf{x}_i) = \frac{\exp\left(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i\right)}{1+\sum_{j=1}^{J-1}\exp\left(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i\right)}, \quad j = 1,\dots,J-1,$$

and the probability for the reference category in the form

$$\pi_{iJ}(\mathbf{x}_i) = 1 - \left[\pi_{i1}(\mathbf{x}_i) + \ldots + \pi_{i(J-1)}(\mathbf{x}_i)\right] = \frac{1}{1 + \sum_{j=1}^{J-1} \exp\left(\alpha_j + \boldsymbol{\beta}_j' \mathbf{x}_i\right)}.$$

Generally, the first or last category is chosen as a reference, but this choice can be arbitrary, depending on the convenience of the researcher (Tang et al., 2012). Furthermore, the effects, $\boldsymbol{\beta}_j$, vary by response category, which implies that the effects of covariates may vary according to the response category being compared to the reference category (Bilder and Loughin, 2014).

Through the maximum likelihood method, the estimation of the parameters of the model (3.1) can be performed. The fit consists of maximizing the probability $\pi_{ij}(\mathbf{x}_i)$ to simultaneously satisfy the $J-1$ equations that specify the model. First, consider the data with individual structure with the observed vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})$ satisfying $\sum_{j=1}^{J} y_{ij} = 1$ and mean $\mathrm{E}(Y_{ij}|\mathbf{x}_i) = \pi_{ij}(\mathbf{x}_i)$, $j = 1, 2, \ldots, J$. Moreover, $y_{iJ} = 1 - \left(\sum_{j=1}^{J-1} y_{ij}\right)$ and $\pi_{iJ}(\mathbf{x}_i) = 1 - \left(\sum_{j=1}^{J-1} \pi_{ij}(\mathbf{x}_i)\right)$, then the logarithm of the likelihood function is given by

$$l = \log \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} \left[\pi_{ij}\left(\mathbf{x}_i\right)\right]^{y_{ij}} \right\} = \log \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J-1} \left[\pi_{ij}\left(\mathbf{x}_i\right)\right]^{y_{ij}} \left[\pi_{iJ}\left(\mathbf{x}_i\right)\right]^{y_{iJ}} \right\}.$$

Using y_{iJ} there is

$$\begin{aligned} l &= \log \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J-1} \left[\pi_{ij}\left(\mathbf{x}_i\right)\right]^{y_{ij}} \left[\pi_{iJ}\left(\mathbf{x}_i\right)\right]^{1-\sum_{j=1}^{J-1} y_{ij}} \right\} \\ &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J-1} y_{ij} \log\left[\pi_{ij}\left(\mathbf{x}_i\right)\right] + \left(1 - \sum_{j=1}^{J-1} y_{ij}\right) \log\left[\pi_{iJ}\left(\mathbf{x}_i\right)\right] \right\} \\ &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J-1} y_{ij} \log\left[\frac{\pi_{ij}(\mathbf{x}_i)}{\pi_{iJ}(\mathbf{x}_i)}\right] + \log\left[\pi_{iJ}\left(\mathbf{x}_i\right)\right] \right\}. \end{aligned}$$

In the last expression the $\log\left[\frac{\pi_{ij}(\mathbf{x}_i)}{\pi_{iJ}(\mathbf{x}_i)}\right]$ is replaced by $\alpha_j + \boldsymbol{\beta}_j' \mathbf{x}_i$ related to the first term and $\pi_{iJ}(\mathbf{x}_i) = \frac{1}{1+\sum_{j=1}^{J-1} \exp(\alpha_j + \boldsymbol{\beta}_j' \mathbf{x}_i)}$ the second term so that

$$\begin{aligned} l &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J-1} y_{ij} \left(\alpha_j + \boldsymbol{\beta}_j' \mathbf{x}_i\right) + \log\left[1/(1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \boldsymbol{\beta}_j' \mathbf{x}_i))\right] \right\}; \\ &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J-1} y_{ij} \left(\alpha_j + \boldsymbol{\beta}_j' \mathbf{x}_i\right) - \log\left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \boldsymbol{\beta}_j' \mathbf{x}_i)\right] \right\}. \end{aligned}$$

To maximize l and obtain the maximum likelihood estimates of the parameters, it is necessary to use iterative methods, which can be done using the Newton-Raphson method, for example (Agresti, 2007).

Now, considering the grouped data where the observed vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})$ satisfies $\sum_{j=1}^{J} y_{ij} = m_i$ with mean $\mathrm{E}(Y_{ij}|\mathbf{x}_i) = m_i \pi_{ij}(\mathbf{x}_i)$, $j = 1, \ldots, J$, there is the likelihood function given by

$$L = \prod_{i=1}^{n} \left\{ \frac{m_i!}{y_{i1}! \ldots y_{iJ}!} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}}\left(\mathbf{x}_i\right) \right\},$$

and the logarithm of the likelihood function by

$$\begin{aligned} l^* &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J} y_{ij} \log\left[\pi_{ij}\left(\mathbf{x}_i\right)\right] + \log\left[\frac{m_i!}{y_{i1}! \ldots y_{iJ}!}\right] \right\} \\ &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J-1} y_{ij} \log\left[\pi_{ij}\left(\mathbf{x}_i\right)\right] + y_{iJ} \log\left[\pi_{iJ}\left(\mathbf{x}_i\right)\right] + \log\left[\frac{m_i!}{y_{i1}! \ldots y_{iJ}!}\right] \right\}. \end{aligned}$$

Replacing $y_{iJ} = m_i - \left(\sum_{j=1}^{J-1} y_{ij}\right)$ in the last expression above there is

$$\begin{aligned} l^* &= \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1} y_{ij}\log\left[\pi_{ij}\left(\mathbf{x}_i\right)\right] + \left(m_i - \sum_{j=1}^{J-1} y_{ij}\right)\log\left[\pi_{iJ}\left(\mathbf{x}_i\right)\right] + \log\left[\frac{m_i!}{y_{i1}!\ldots y_{iJ}!}\right]\right\} \\ &= \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1} y_{ij}\log\left[\pi_{ij}\left(\mathbf{x}_i\right)\right] + m_i\log\left[\pi_{iJ}\left(\mathbf{x}_i\right)\right] - \sum_{j=1}^{J-1} y_{ij}\log\left[\pi_{iJ}\left(\mathbf{x}_i\right)\right] + \log\left[\frac{m_i!}{y_{i1}!\ldots y_{iJ}!}\right]\right\} \\ &= \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1} y_{ij}\log\left[\frac{\pi_{ij}(\mathbf{x}_i)}{\pi_{iJ}(\mathbf{x}_i)}\right] + m_i\log\left[\pi_{iJ}\left(\mathbf{x}_i\right)\right] + \log\left[\frac{m_i!}{y_{i1}!\ldots y_{iJ}!}\right]\right\}. \end{aligned}$$

Finally, $\log\left[\frac{\pi_{ij}(\mathbf{x}_i)}{\pi_{iJ}(\mathbf{x}_i)}\right]$ is replaced by $\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i$ and $\pi_{iJ}(\mathbf{x}_i) = \frac{1}{1+\sum_{j=1}^{J-1}\exp(\alpha_j+\boldsymbol{\beta}_j'\mathbf{x}_i)}$ in the expression above such that

$$\begin{aligned} l^* &= \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1} y_{ij}(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i) + m_i\log\left[1/(1+\sum_{j=1}^{J-1}\exp(\alpha_j+\boldsymbol{\beta}_j'\mathbf{x}_i))\right] + \log\left[\frac{m_i!}{y_{i1}!\ldots y_{iJ}!}\right]\right\} \\ &= \sum_{i=1}^{n}\left\{\sum_{j=1}^{J-1} y_{ij}(\alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i) - m_i\log\left[1+\sum_{j=1}^{J-1}\exp(\alpha_j+\boldsymbol{\beta}_j'\mathbf{x}_i)\right] + \log\left[\frac{m_i!}{y_{i1}!\ldots y_{iJ}!}\right]\right\}. \end{aligned}$$

An iterative method such as the Newton-Raphson method can be used to maximize l^* and obtain the maximum likelihood estimates (Tutz, 2011).

3.2.4 Residuals associated with models for nominal categorized data

According to Nobre and Singer (2007), it is essential to verify the robustness of the estimates obtained from the fit, based on the statistical model under analysis, to describe the observed data well, leading the researcher to reliable inferences and predictions. An important step performed for the diagnostics of a model is the analysis of residuals. Residuals are used to validate the assumptions of a model and detect outliers or influential points. In this way, they can be used as tools that constitute the selection of a model.

3.2.4.1 Residuals for individual data

The ordinary residual, which measures the deviations between the observed values and the predicted probabilities, for the model (3.1) is a vector with dimension $J\times 1$ per individual i, $i = 1, 2, \ldots, n$, given by (Reiter and Kohnen, 2005)

$$\hat{\mathbf{r}}_i = \mathbf{y}_i - \hat{\boldsymbol{\pi}}_i = \left(y_{i1} - \hat{\pi}_{i1}, y_{i2} - \hat{\pi}_{i2}, \ldots, y_{iJ} - \hat{\pi}_{iJ}\right)',$$

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})'$ is the observed vector with $y_{ij} = 1$ if the response of individual i belongs to the category j and $y_{ij} = 0$, otherwise and $\boldsymbol{\pi}_i = (\hat{\pi}_{i1}, \hat{\pi}_{i2}, \ldots, \hat{\pi}_{iJ})'$ is the vector of predicted probabilities. This residual does not follow a normal distribution, and when used in diagnostic plots, it may not be informative, generating great visual difficulties and interpretation.

The Pearson and deviance residuals for the model (3.1)) are given, respectively, by the vectors $r_i^P = \left[r_{i1}^P, r_{i2}^P, \ldots, r_{iJ}^P\right]'$ and $r_i^D = \left[r_{i1}^D, r_{i2}^D, \ldots, r_{iJ}^D\right]'$, whose elements are obtained by (Cheng et al., 2021)

$$r_{ij}^P = \frac{(y_{ij} - \hat{\pi}_{ij})}{\sqrt{\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}}$$

and

$$r_{ij}^D = I\left(y_{ij} - \hat{\pi}_{ij}\right)\sqrt{2\left[(y_{ij}-1)\log\left(1-\hat{\pi}_{ij}\right) - y_{ij}\log\left(\hat{\pi}_{ij}\right)\right]}$$

where $j = 1, 2, \ldots, J$, and with the indicator function $I\left(y_{ij} - \hat{\pi}_{ij}\right) = 1$ if $y_{ij} - \hat{\pi}_{ij} > 0$ and -1, otherwise. These definitions are extensions of the Pearson and deviance residuals used to assess the fit of the logistic regression model.

The residual proposed by Cheng et al. (2021) is based on the methodology presented by Liu and Zhang (2018) in defining a continuous variable to replace the original and the surrogate residuals obtained from this new variable. Let a continuous random vector per individual, $\mathbf{U}_i$, J-dimensional that corresponds to a deterministic part (linear predictor structure of the model (3.1)) and a random part, with elements given by

$$U_{ij} = \alpha_j + \boldsymbol{\beta}_j' \mathbf{x}_i + \varepsilon_{ij},$$

where $\varepsilon_{i1}, \ldots, \varepsilon_{iJ}$ are mutually independent with standard Gumbel distribution, $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, J$. The distribution of the random part depends on the link function used in the model, being assumed as another one for a different link function.

Consider now the joint distribution between the original response variable and the continuous variable established by $Y_{ij} = 1$, the response of individual i belongs to category j, if and only if $U_{ij} > U_{ij'},\ \forall j \neq j'$, with $j, j' = 1, 2, \ldots, J$. The continuous random vector $\mathbf{S}_i$ is defined following the conditional distribution of $\mathbf{U}_i = (U_{i1}, \ldots, U_{iJ})$ given the observed vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})'$, more specifically, $\mathbf{S}_i \sim \mathbf{U}_i | (\mathbf{y}_i, \mathbf{x}_i)$ using the region $U_{ij} > U_{ij'}$. Then, the vector of surrogate residuals is expressed as

$$\mathbf{R}_i = \mathbf{S}_i - \mathrm{E}_0(\mathbf{S}_i | \mathbf{x}_i) = \mathbf{S}_i - \boldsymbol{\beta}_j \mathbf{x}_i' - \mathrm{E}(\boldsymbol{\varepsilon}_i) \tag{3.2}$$

where E(.) denotes mean and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})'$ follows the J-dimensional standard Gumbel distribution. If the model (3.1) is specified correctly, the random vector $\mathbf{S}_i$ follows the same distribution as $\mathbf{U}_i$ and the residual, which is also a continuous vector, has the following properties (Cheng et al., 2021):

i) Zero mean vector: $\mathrm{E}(\mathbf{R}_i | \mathbf{x}_i) = \mathbf{0}$.

ii) Constant variance matrix: $\mathrm{Var}(\mathbf{R}_i | \mathbf{x}_i) = \mathrm{Var}(\boldsymbol{\varepsilon}_i)$.

iii) The random vector $\mathbf{R}_i$, independent from $\mathbf{x}_i$, follows the J-dimensional standard Gumbel distribution.

For small samples, the authors used the bootstrap method presented in Liu and Zhang (2018) to obtain the empirical distributions of residuals by resampling. Furthermore, the vector of surrogate residuals was normalized since the standard Gumbel distribution is positive skew and can lead to a biased inference if the skewness is not considered.

3.2.4.2 Residuals for grouped data

The J-dimensional ordinary residual vector for the model (3.1) per experimental unit i, $i = 1, 2, \ldots, n$, is obtained by (Tutz, 2011)

$$\begin{aligned} \hat{\mathbf{r}}_i &= \frac{\mathbf{y}_i - m_i \times \hat{\boldsymbol{\pi}}_i}{m_i} \\ &= \frac{1}{m_i}(y_{i1} - m_i\hat{\pi}_{i1}, y_{i2} - m_i\hat{\pi}_{i2}, \ldots, y_{iJ} - m_i\hat{\pi}_{iJ})', \end{aligned}$$

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})'$ is the observed vector with the counts in the j-th category, y_{ij}, for the experimental unit i and fixed size m_i such that $\sum_{j=1}^{J} y_{ij} = m_i$ and $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i1}, \hat{\pi}_{i2}, \ldots, \hat{\pi}_{iJ})'$ is the vector of predicted probabilities.

The vector of Pearson residuals J-dimensional is given by $r_i^P = \left[r_{i1}^P, r_{i2}^P, \ldots, r_{iJ}^P\right]'$ with elements obtained as follows (Tutz, 2011)

$$r_{ij}^P = \frac{(y_{ij} - m_i\hat{\pi}_{ij})}{\sqrt{m_i\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}}$$

where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, J$.

3.3 Randomized quantile residual

The quantile residual was proposed by Dunn and Smyth (1996) for situations with a continuous variable and extended to situations with a discrete variable. For a continuous response, y_i, the quantile residual is defined by

$$r_i^Q = \Phi^{-1}\left\{F(y_i;\hat{\theta}_i,\hat{\phi})\right\}, \;\; i = 1, \ldots, n,$$

where Φ^{-1} is the cumulative distribution function (CDF) of the standard normal distribution, $F(y_i;\hat{\theta}_i,\hat{\phi})$ is the CDF associated with response variable, $\hat{\theta}_i$ is the maximum likelihood estimate of the parameter θ_i and the dispersion parameter is $\hat{\phi}$.

On the other hand, if the response y_i is discrete, a more general definition of randomized quantile residuals is necessary. The idea is to introduce randomization through a uniform random component in the CDF for each individual. Thus, the randomized quantile residual is obtained by

$$r_i^Q = \Phi^{-1}\left\{F(u_i)\right\}, \;\; i = 1, \ldots, n,$$

where u_i represents a uniform random variable between $a_i = \lim_{y\to y_i} F(y;\hat{\theta}_i,\hat{\phi})$ and $b_i = F(y_i;\hat{\theta}_i,\hat{\phi})$. Then, these residuals also follow an approximately normal distribution.

The quantile residuals can be used for a broad class of regression models, being easy to compute in statistical software. However, they have received little attention in the literature as model diagnostic tools until recently. For example, Klar and Meintanis (2012) used the standardized quantile residuals in goodness-of-fit tests for generalized linear models with inverse Gaussian and gamma variables. In their study, these residuals showed to follow an approximately standard normal distribution. Pereira (2019) investigated the performance of the quantile residual for diagnostics of the beta regression model and demonstrated that this residual is better approximated by the standard normal distribution than other residuals in several scenarios. Furthermore, Feng et al. (2020) used the standardized randomized quantile residuals to examine the fit of models to count data, including the zero-inflated model, and obtained satisfactory results.

One issue that may raise concern is the fluctuation of the randomized quantile residuals due to the randomization introduced to obtain the continuous residuals (Feng et al., 2020). Dunn and Smyth (1996) recommended producing the residuals multiple times, ensuring that discrepancies do not occur due to randomization.

3.4 Distances

Consider having n individuals denoted by the random vectors $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iq})'$, $i = 1, 2, \ldots, n$. Each individual is represented by a point in q-dimensional space, with each dimension representing a variable (Sharma, 1996). As an example, we can mention the data presented by Johnson and Wichern (2007) about 25 lizards Cophosaurus texanus, in which measurements were obtained for each individual regarding the three variables: weight (or mass), in grams, snout-vent length (SVL) and hind limb span (HLS), both in millimeters. The scatterplot in $\mathbb{R}^3$ for the variables in this example can be seen in Figure 3.1.

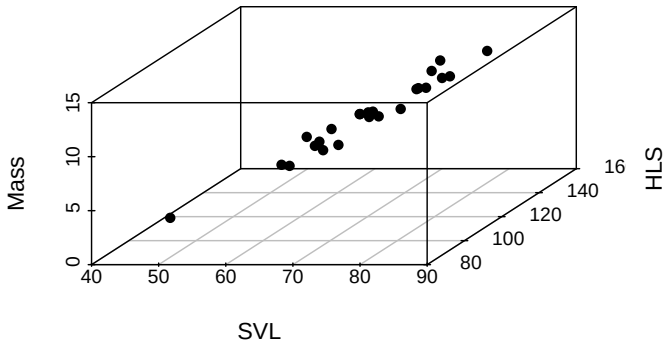

Figure 3.1. Scatter plot of lizard data in three-dimensional space (HLS, mass, SVL) presented in Johnson and Wichern (2007) p. 19.

The distance measures can quantify how far two individuals are by a scalar, measuring their proximity. The Euclidean and Mahalanobis distances are widely known in multivariate statistics and can be calculated in the original space of the response variable (Maesschalck et al., 2000). While the Euclidean distance is simple to calculate and interpret, the Mahalanobis distance has an invariant scale and takes into account the correlation in the data since it is calculated using the inverse of the covariance matrix of the set of interest. In addition, they are widely used in various classification techniques, in cluster analysis as well as to detect outliers, especially in the context of linear regression models (see Zelterman (2015) and Kannan and Manoj (2015)).

To measure the distance between an individual i and an individual t, the Euclidean distance is defined by

$$d_{it}^{E} = \sqrt{(\mathbf{z}_i - \mathbf{z}_t)'(\mathbf{z}_i - \mathbf{z}_t)} = \sqrt{\sum_{k=1}^{q} (z_{ik} - z_{tk})},$$

where $z_{\cdot k}$ is the k-th variable, with $k = 1, 2, \ldots, q$, and $i, t = 1, 2, \ldots, n$. According to Zelterman (2015), this measure is the most popular to calculate the distance between individuals in q-dimensional space. On the other hand, if the individuals are correlated, the covariance or correlation between them must be considered when calculating the distance (Sharma, 1996). In this case, one can obtain the Mahalanobis distance expressed by

$$d_{it}^{2} = (\mathbf{z}_i - \mathbf{z}_t)'\mathbf{C}^{-1}(\mathbf{z}_i - \mathbf{z}_t),$$

where $\mathbf{C}^{-1}$ is the covariance matrix with dimension $q \times q$. In the case where $\mathbf{C} = \mathbf{I}$, with $\mathbf{I}$ representing the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If $\mathbf{C}$ is a diagonal matrix, then the distance results in the standardized Euclidean distance (Johnson and Wichern, 2007).

The Euclidean distance results in quicker calculations than the Mahalanobis distance, but considering the correlation is an important factor given its frequent observation in practice (Ghorbani, 2019). However, Maesschalck et al. (2000) reported that some issues must be observed in the Mahalanobis distance, such as a large number of variables that can lead to a singular covariance matrix and the sample size that must be greater than the number of variables.

3.5 Methods

This section describes the methodological procedures for analyzing residuals and diagnostics associated with the generalized logit models for nominal polytomous data, distinguishing the experimental designs with individual and grouped structures. The randomized quantile residual is described for individual data. In this way, the normality of this residual can be assessed in simulation studies to validate the generalized logit model, an application not found in the literature. In addition, a new methodology is presented to reduce the dimension of ordinary residuals associated with grouped data through Euclidean and Mahalanobis distances as a diagnostic tool to detect outliers. In this context, the steps to carry out are:

i) Fit the generalized logit model presented in section 3.2.3 with different linear predictors in motivation and simulation studies.

ii) Carry out the selection of variables that will form the linear predictor of the model through the likelihood ratio test. Considering two nested models, M_0 and M_1, given by

$$M_0 : \log\left[\frac{\pi_{ij}(\mathbf{x}_i)}{\pi_{iJ}(\mathbf{x}_i)}\right] = \alpha_j \quad \text{and} \quad M_1 : \log\left[\frac{\pi_{ij}(\mathbf{x}_i)}{\pi_{iJ}(\mathbf{x}_i)}\right] = \alpha_j + \boldsymbol{\beta}_j'\mathbf{x}_i, \quad j = 1, \ldots, J-1.$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is the vector of p covariates, $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})'$ is the vector of regression parameters, and α_j is the intercept.

The hypotheses to be tested are

$$\begin{cases} H_0 : \boldsymbol{\beta}_j = \mathbf{0}, \ \forall j = 1, \ldots, J-1 \\ H_1 : \exists\, j \,|\, \boldsymbol{\beta}_j \neq \mathbf{0} \end{cases}$$

and the test statistic

$$\Lambda = -2\log\left[\frac{L_{H_0}}{L_{H_1}}\right] = 2\log(L_{H_1}) - 2\log(L_{H_0}) \sim \chi^2_{(m)},$$

where L_{H_0} is the likelihood function associated with the model M_0, L_{H_1} is the likelihood function associated with the model M_1 (with the variable(s) under investigation) and m is the difference in parameters between the two models. The null hypothesis is rejected at the 5% significance level when $\Lambda > \chi^2_{(m)}$.

The Akaike Information Criterion (AIC) (Akaike, 1974) can also be used in the model selection stage, discriminating the models by the different linear predictors. The model with the smallest distance from the probabilistic process that generated the data will be the best, that is, the linear predictor with the lowest AIC will be the one indicated. Thus, the AIC is defined by

$$\text{AIC} = -2\hat{l}_p + 2p,$$

where $\hat{l}_p$ is the logarithm of the likelihood function and p is the number of model parameters under search.

The $100(1-\alpha)\%$ confidence intervals (CIs) can be constructed for the parameters, and these are given by (Agresti, 2002)

$$\hat{\alpha}_j \pm z_{(1-\alpha/2)} \times \sqrt{\hat{\text{Var}}(\hat{\alpha}_j)}$$

and

$$\hat{\beta}_{j1} \pm z_{(1-\alpha/2)} \times \sqrt{\hat{\text{Var}}(\hat{\beta}_{j1})}, \ \ldots, \ \hat{\beta}_{jp} \pm z_{(1-\alpha/2)} \times \sqrt{\hat{\text{Var}}(\hat{\beta}_{jp})}$$

where $j = 1, 2, \ldots, J-1$, $\alpha = 0,05$ is the significance level, and z is the quantile of the normal distribution.

iii) For diagnostics of the generalized logit model, the methodologies present in the residual analysis are:

1) Obtain the standardized randomized quantile residuals for the individual data considering the cumulative distribution function (CDF), $F(\mathbf{y}_i;\hat{\boldsymbol{\pi}}_i,\hat{\phi})$, for the response vector $\mathbf{y}_i$ given the vector $\mathbf{x}_i$, $i=1,2,\ldots,n$. The CDF for multinomial distribution was presented by Levin (1981) using its representation as a conditional distribution of independent Poisson random variables given a fixed sum. For a small number of categories, J, the multinomial CDF is as easy to compute (exactly) as the convolution of J truncated Poisson random variables.

Let the estimated parameter given by $\hat{\boldsymbol{\pi}}_i=(\hat{\pi}_{i1}(x_i),\hat{\pi}_{i2}(x_i),\ldots,\hat{\pi}_{iJ}(x_i))'$, and ϕ is the dispersion parameter that does not depend on $\mathbf{x}_i$. Consider the probability mass function $f(\mathbf{y}_i;\hat{\boldsymbol{\pi}}_i,\hat{\phi})$, corresponding the response of individual i in category j, $y_{ij}=1$, and $y_{ij}=0$ otherwise. Then, the estimated cumulative distribution function for individual i is

$$F^*(\mathbf{y}_i,u_i;\hat{\boldsymbol{\pi}}_i,\hat{\phi})=F(\mathbf{y}_i-;\hat{\boldsymbol{\pi}}_i,\hat{\phi})+u_i\times f(\mathbf{y}_i;\hat{\boldsymbol{\pi}}_i,\hat{\phi}),$$

where $F(\mathbf{y}_i-;\hat{\boldsymbol{\pi}}_i,\hat{\phi})$ is the CDF for the vector $\mathbf{y}_i-$ that receives zero in the place of observed category by the vector $\mathbf{y}_i$ and one in the other categories. Also, u_i is a random variable with uniform distribution of parameters $(0,1)$, and using $\hat{\phi}=1$.

The randomized quantile residual for a polytomous response $\mathbf{y}_i$ is given by

$$r_i^Q=\Phi^{-1}[F^*(\mathbf{y}_i,u_i;\hat{\boldsymbol{\pi}}_i,\hat{\phi})],$$

where Φ^{-1} is the quantile function of the standard normal distribution. Thus, this resultant residual is a single value for each i, approximating a normal distribution if the model is specified correctly.

In this work, the randomized quantile residuals were standardized, being obtained by

$$r_i^S=\frac{r_i^Q-\bar{r}^Q}{S_{r^Q}},$$

where $\bar{r}^Q=n^{-1}\sum_{i=1}^{n}r_i^Q$ and $S_{r^Q}^2=(n-1)^{-1}\sum_{i=1}^{n}\left(r_i^Q-\bar{r}^Q\right)^2$ are respectively the mean and the variance of residuals.

2) Reduce the vector dimension of ordinary residuals associated with the grouped data through Euclidean and Mahalanobis distances. Consider the vector of ordinary residuals for the experimental unit i, described in section 3.2.4.2, with the zero mean vector $\mathrm{E}(\mathbf{r}_i|\mathbf{x}_i)=\mathbf{0}$ of dimension J under the assumption that the model is specified correctly. Then, the expressions using Euclidean and Mahalanobis distances in multinomial regression are expressed, respectively, by

$$d_i^M=\sqrt{(\mathbf{r}_i-\mathbf{0})'(\mathbf{r}_i-\mathbf{0})}=\sqrt{\sum_{j=1}^{J}r_{ij}^2},\ \ j=1,2,\ldots,J$$

and

$$d_{Mi}^2=(\mathbf{r}_i-\mathbf{0})'\mathbf{C}_i^{-1}(\mathbf{r}_i-\mathbf{0})=\mathbf{r}_i'\mathbf{C}_i^{-1}\mathbf{r}_i,$$

where $\mathbf{C}_i$, with dimension $J\times J$, is the covariance matrix of the residuals for the experimental unit i, $i=1,\ldots,n$, i.e.,

$$\mathbf{C}_i=\begin{bmatrix}\mathrm{Var}(r_{i1}) & \mathrm{Cov}(r_{i1},r_{i2}) & \ldots & \mathrm{Cov}(r_{i1},r_{iJ})\\ \mathrm{Cov}(r_{i2},r_{i1}) & \mathrm{Var}(r_{i2}) & \ldots & \mathrm{Cov}(r_{i2},r_{iJ})\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{Cov}(r_{iJ},r_{i1}) & \mathrm{Cov}(r_{iJ},r_{i2}) & \ldots & \mathrm{Var}(r_{iJ})\end{bmatrix}.$$

iv) Once the randomized quantile residuals and distance measures are defined, formal (tests) and informal (plots) techniques are employed for diagnostics.

A powerful and widely known test for detecting deviations from normality due to asymmetry or kurtosis (or both) is the test of Shapiro and Wilk (1965). Consider the standardized randomized quantile residuals described in the step (iii) of this section given by r_i^S, $i = 1, \ldots, n$, and sorted by $r_{(1)}^S < r_{(2)}^S < \ldots < r_{(n)}^S$. The hypotheses to be tested are

$$\begin{cases} H_0 : r_i^S \sim N(0,1) \\ H_1 : r_i^S \text{ does not follow } N(0,1) \end{cases}$$

with test statistic

$$SW = \frac{\left(\sum_{i=1}^{n} a_i r_{(i)}^S\right)^2}{\sum_{i=1}^{n} \left(r_i^S - \bar{r}_i^S\right)^2}$$

where $\bar{r}_i^S$ is the residual mean and $(a_1, a_2, \ldots, a_n) = \frac{\mathbf{m}'\mathbf{V}^{-1}}{(\mathbf{m}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m})^{1/2}}$, with $\mathbf{m} = (m_1, m_2, \ldots, m_n)'$ denoting the vector of expected values of the order statistics of the standard normal distribution and $\mathbf{V}$ is the covariance matrix $n \times n$. Initially, the test was restricted to a sample size up to 50, but Royston (1995) provided an algorithm with an improved approximation to the weights a_i, $i = 1, 2, \ldots, n$, which can be used for any sample with size between 3 and 5000.

Considering informal techniques, one can first visualize the distribution of residuals through a histogram, comparing its shape with that of the normal distribution. In the plot of residuals versus fitted values, it is possible to observe the existence of variance heterogeneity or the presence of outliers. The expected pattern in this plot is the zero-centered distribution of residuals with constant amplitude (Faraway, 2016).

In the half-normal plot, the performance of the residuals and the distance measures show whether the observed data are a plausible realization of the fitted model. The absolute values of a given diagnostic measure (residuals or distances) are compared in relation to the expected order statistics of the half-normal distribution obtained by

$$\Phi^{-1}\left[\frac{(i+n-1/8)}{2n+1/2}\right],$$

where Φ^{-1} is the standard normal distribution function, with $i = 1, \ldots, n$ and n the sample size. Atkinson (1985) proposed adding simulated envelopes obtained through computer simulations to the plot to simplify the interpretation. The steps to build the envelope are(Moral et al., 2017): 1) Fit the model and obtain the sorted absolute values of a diagnostic measure, $d_{(i)}$; 2) Using the fitted model, simulate 99 samples for the response variable; 3) Fit the model for each simulated sample, obtaining the absolute and sorted values of the diagnostic measure, $d_{t(i)}^*$, $t = 1, \ldots, 99$ and $i = 1, \ldots, n$; 4) For each set, calculate the percentiles $2,5\%, 50\%, 97,5\%$; 5) Plot these percentiles and the $d_{(i)}$ of the original sample against the order statistics of the half-normal distribution. Then, a considerable proportion of points outside the envelope indicates that the model is not suitable for analyzing the data, being a satisfactory fit when the number of points outside the envelope is equal to or less than 5%.

v) Finally, simulation studies are developed to evaluate the performance of the described procedures.

3.6 Simulation studies

A simulation study was developed to evaluate the performance of standardized quantile residuals and distance measures in diagnostic techniques (formal and informal) in relation to the correct

specification of the linear predictor structure. The generalized logit models with three response categories for individual and grouped structures and two linear predictor structures define the simulation scenarios. The first mean structure has the effect of a continuous covariate. In contrast, in the second, the model has the effect of two covariates, one being continuous and the other a factor.

3.6.1 Models and scenarios

Eight scenarios are considered, obtained by combining two linear predictor structures, nominal polytomous data in individual structure with sample size $n = 100$ and nominal data in grouped structure with group dimensions ($m = 5, 10$ e 15) and sample size $n = 50$. For each scenario, 1000 data sets were simulated. In the scenario of the linear predictor structure of model 1 (continuous covariate), the response variable was simulated from the generalized logit model expressed by

$$\log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = \alpha_j + \beta_j x_i, \quad j = 2, 3,$$

for which it was assumed that $X \sim N(0, 1)$; $\alpha_2 = 1,38$ and $\alpha_3 = 3,51$ for the intercepts; $\beta_2 = -2,7$ and $\beta_3 = -5,11$ for the regression parameters, $i = 1, 2, \ldots, n$. The first category was set as the reference. The model probabilities are expressed by

$$\pi_{i2}(x_i) = \frac{\exp(\alpha_2 + \beta_2 x_i)}{1 + \exp(\alpha_2 + \beta_2 x_i) + \exp(\alpha_3 + \beta_3 x_i)},$$

$$\pi_{i3}(x_i) = \frac{\exp(\alpha_3 + \beta_3 x_i)}{1 + \exp(\alpha_2 + \beta_2 x_i) + \exp(\alpha_3 + \beta_3 x_i)}$$

and

$$\pi_{i1}(x_i) = 1 - \pi_{i2}(x_i) - \pi_{i3}(x_i).$$

In the scenario with the linear predictor structure of model 2 (continuous and factor covariates), the response variable was simulated using a generalized logit model given by

$$\log\left(\frac{\pi_{ij}}{\pi_{i1}}\right) = \alpha_j + \beta_{j1} x_{i1} + \beta_{j2} x_{i2}, \quad j = 2, 3, \tag{3.3}$$

for which the first category was defined as the reference, $\alpha_2 = 1,38$ and $\alpha_3 = 3,51$ for the intercepts; $\beta_{21} = -2,7$, $\beta_{22} = 1,35$, $\beta_{31} = -5,11$ and $\beta_{32} = 2,49$ for the regression parameters , $X_{i1} \sim N(0, 1)$ is the continuous covariate and X_{i2} is a factor with 2 levels, "control" and "treatment" considered in the model through the dummy variable, respectively. Furthermore, the probabilities for the model are

$$\pi_{i2}(x_{i1}, x_{i2}) = \frac{\exp(\alpha_2 + \beta_{21} x_{i1} + \beta_{22} x_{i2})}{1 + \exp(\alpha_2 + \beta_{21} x_{i1} + \beta_{22} x_{i2}) + \exp(\alpha_3 + \beta_{31} x_{i1} + \beta_{32} x_{i2})},$$

$$\pi_{i3}(x_{i1}, x_{i2}) = \frac{\exp(\alpha_3 + \beta_{31} x_{i1} + \beta_{32} x_{i2})}{1 + \exp(\alpha_2 + \beta_{21} x_{i1} + \beta_{22} x_{i2}) + \exp(\alpha_3 + \beta_{31} x_{i1} + \beta_{32} x_{i2})}$$

and

$$\pi_{i1}(x_{i1}, x_{i2}) = 1 - \pi_{i2}(x_{i1}, x_{i2}) - \pi_{i3}(x_{i1}, x_{i2}).$$

In cases where individual data are considered, that is, $y_{ij} = 1$ if the outcome of individual i belongs the category j, $j = 1, 2, 3$, and $y_{ij} = 0$, otherwise, $i = 1, 2, \ldots, 100$. The randomized quantile residual of $\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3})'$ is expressed by

$$r_i^Q = \Phi^{-1}[F^*(\mathbf{y}_i, u_i; \hat{\boldsymbol{\pi}}_i, \hat{\phi})],$$

where $F^*(\mathbf{y}_i, u_i; \hat{\boldsymbol{\pi}}_i, \hat{\phi})$ is the estimated cumulative distribution function for the observed response, with estimated mean $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i1}, \hat{\pi}_{i2}, \hat{\pi}_{i3})'$, $u_i \sim \mathrm{U}(0,1)$ and $\hat{\phi} = 1$. In addition, Φ^{-1} represents the quantile function of the normal standard distribution. Thus, standardized randomized quantile residuals are obtained from

$$r_i^S = \frac{r_i^Q - \bar{r}^Q}{S_{r^Q}},$$

where $\bar{r}^Q$ and $S_{r^Q}^2$ are the mean and variance of the residuals, respectively.

Now, in the case of grouped data, the vector of ordinary residuals $\hat{\mathbf{r}}_i = (\hat{r}_{i1}, \hat{r}_{i2}, \hat{r}_{i3})'$, $i = 1, 2, \ldots, 50$, is given by

$$\hat{\mathbf{r}}_i = \frac{\mathbf{y}_i - m_i\hat{\boldsymbol{\pi}}_i}{m_i} = \left(\frac{y_{i1} - m_i \times \hat{\pi}_{i1}}{m_i}, \frac{y_{i2} - m_i \times \hat{\pi}_{i2}}{m_i}, \frac{y_{i3} - m_i \times \hat{\pi}_{i3}}{m_i}\right)'$$

where $\mathbf{y} = (y_1, y_2, y_3)'$ represents the vector with the counts in relation to the categories observed in the group, with $m_i = y_{i1} + y_{i2} + y_{i3}$, $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i1}, \hat{\pi}_{i2}, \hat{\pi}_{i3})'$ is the vector of predicted probabilities. Then, the Euclidean and Mahalanobis distances are defined, respectively, by

$$d_i^M = \sqrt{\hat{\mathbf{r}}_i'\hat{\mathbf{r}}_i}$$

and

$$d_{Mi}^2 = \hat{\mathbf{r}}_i'\hat{\mathbf{C}}_i^{-1}\hat{\mathbf{r}}_i,$$

where $\hat{\mathbf{C}}_i$, with dimension 3×3, is the covariance matrix of residuals, that is,

$$\hat{\mathbf{C}}_i = \begin{bmatrix} \hat{\mathrm{Var}}(\hat{r}_{i1}) & \hat{\mathrm{Cov}}(\hat{r}_{i1}, \hat{r}_{i2}) & \hat{\mathrm{Cov}}(\hat{r}_{i1}, \hat{r}_{i3}) \\ \hat{\mathrm{Cov}}(\hat{r}_{i2}, \hat{r}_{i1}) & \hat{\mathrm{Var}}(\hat{r}_{i2}) & \hat{\mathrm{Cov}}(\hat{r}_{i2}, \hat{r}_{i3}) \\ \hat{\mathrm{Cov}}(\hat{r}_{i3}, \hat{r}_{i1}) & \hat{\mathrm{Cov}}(\hat{r}_{i3}, \hat{r}_{i2}) & \hat{\mathrm{Var}}(\hat{r}_{i3}) \end{bmatrix}.$$

Finally, to fit the models, the multinom(.) function was used of the nnet package (Ripley and Venables, 2016). The dmultinom(.) function from the stats package and the pmultinom(.) function from the pmultinom package (Davis, 2018) were used to create the function to obtain the standardized randomized quantile residuals for the proposed scenarios available in the appendix. The calculation of Euclidean and Mahalanobis distances were performed, respectively, through the dist(.) and mahalanobis(.) functions of the stats package (R Core Team, 2020). For diagnostic techniques, the p-value of the Shapiro-Wilk test was obtained by shapiro.test(.) function of the stats package (R Core Team, 2020) and the half-normal plot with simulated envelope was obtained by hnp(.) function of the hnp package (Moral et al., 2017) implemented for a series of generalized models and extensions. All functions are available in R software (R Core Team, 2020).

3.6.2 Results

This section presents the results of the scenarios considered in this work. First, the normality of the randomized quantile residuals for the correct and null models was evaluated using the Shapiro-Wilk test. The p-values obtained in the 1000 simulations were plotted on the histogram to visually understand the distribution of these values. By doing this, it was possible to distinguish the variation in the distribution of p-values about their measurements, such as amplitude, symmetry, and position around the central value. Thereafter, the values obtained from distances were evaluated in the half-normal plot with a simulated envelope. Then, boxplots were used to explore the distribution of the number of points outside the envelope obtained in the simulations using this approach. This way, it was possible to compare the correct and null models regarding the number of points outside the envelope (median, dispersion, and outliers) and compare the distances.

3.6.2.1 Simulation results for randomized quantile residual

Firstly, it is considered the results referring to model 1 compared to the null model (intercept effect) for the individual data based on the 1000 simulations. The histograms of the p-values of the Shapiro-Wilk test under model 1 and the null model are presented in Figure 3.2. The p-values under the null model, Figure 3.2(a), show a positively skewed distribution with a mode close to zero. Conversely, the p-values under model 1, Figure 3.2(b), are practically a uniform pattern. Thus, the normality of residuals was rejected by the test (p-value$< 0,05$) in most simulations considering the null model. This fact did not occur for the correct model being possible to conclude that the standardized randomized quantile residuals performed well in this simulation scenario.

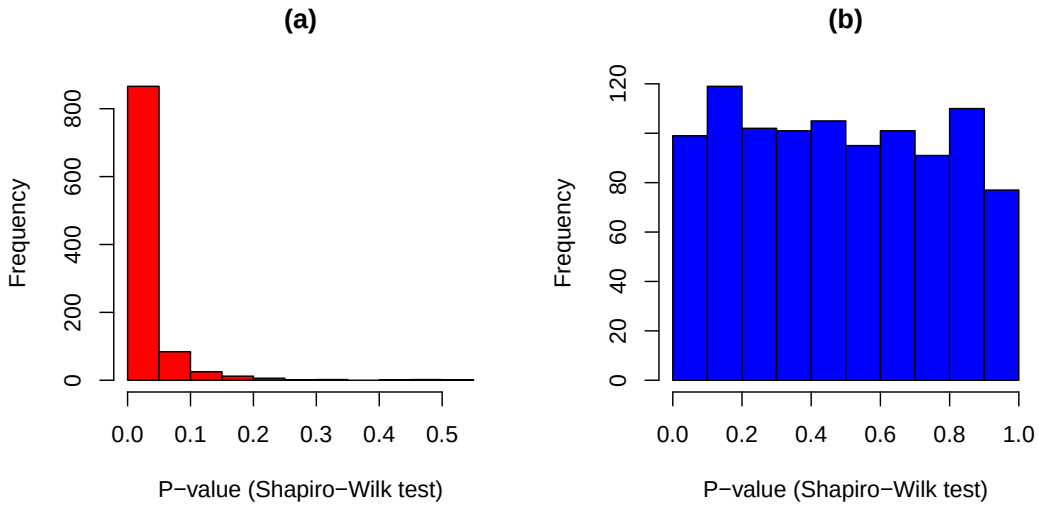

Figure 3.2. Histograms of p-values (Shapiro-Wilk test) for the standardized randomized quantile residuals in the 1000 simulations under (a) null model (intercept effect) (b) model 1 (continuous covariate).

The Shapiro-Wilk test rejected the hypothesis of normality of the residuals in only 40 of 1000 simulations with the p-value mean of $0,483$ for model 1. On the other hand, the Shapiro-Wilk test rejected the hypothesis of residual normality for the null model in 866 simulations with a mean of p-values equal to $0,025$.

Now, it is presented the performance results of the standardized randomized quantile residuals for model 2 (continuous and factor covariates) compared to the null model (intercept effect only) for the individual data. The histograms in 3.3 refer to the p-values resulting from the Shapiro-Wilk test to examine the normality of residuals under the models. The p-values of the Shapiro-Wilk test for model 2 present a relatively uniform distribution, Figure 3.3(b), while the p-values for the null model are clustered close to zero, Figure 3.2(a).

When fitting the null model, the Shapiro-Wilk test rejected the hypothesis of normality of the residuals in 927 of 1000 simulations and presented a mean of p-values equal to $0,015$. In comparison, the Shapiro-Wilk test rejected the normality assumption of the residuals in only 57 simulations for model 2, with a mean of p-values equal to $0,462$.

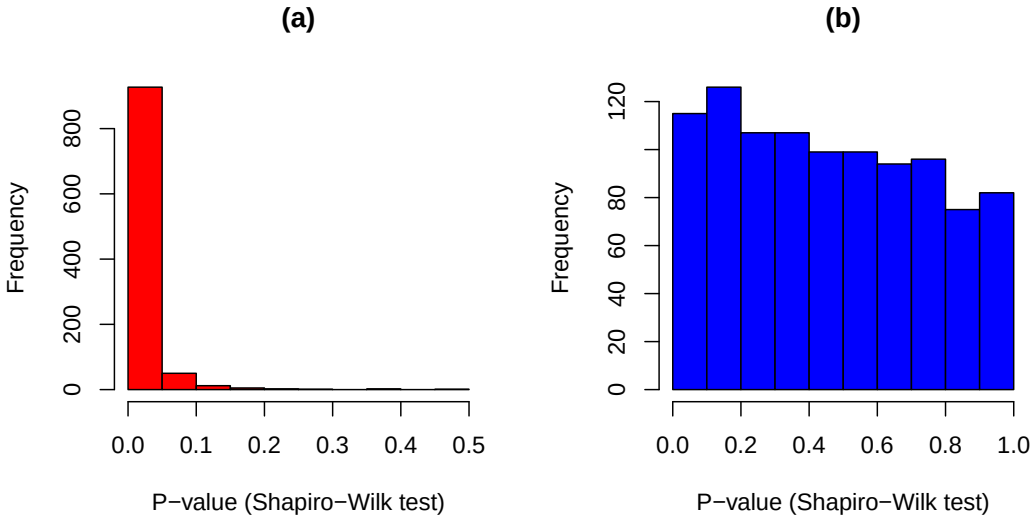

Figure 3.3. Histograms of p-values (Shapiro-Wilk test) for the standardized randomized quantile residuals in the 1000 simulations under (a) null model (intercept effect) (b) model 2 (continuous and factor covariates).

3.6.2.2 Simulation results for distance measures

The results for the grouped data in which $m = 15$ were chosen to display since the values were similar for $m = 5$ and $m = 10$, indicating that the group dimension did not represent a source of variation in the Euclidean and Mahalanobis distance measures, particularly in this study.

It was possible to distinguish model 1 from the null model in the developed studies by the half-normal plot with simulated envelopes for the values of Euclidean and Mahalanobis distances considering a grouped structure of the data. The boxplots, Figure 3.4, refer to the numbers of points outside the envelope obtained in the 1000 simulations for each model using the distance measures.

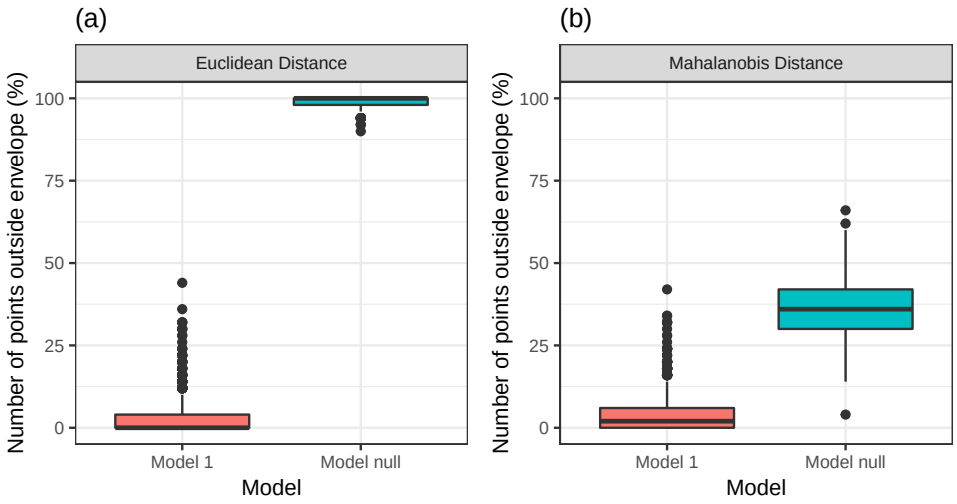

Figure 3.4. Boxplots of the number of points outside the simulated envelope of model 1 (continuous covariate) and null model (intercept effect) referring to (a) Euclidean and (b) Mahalanobis distances.

The median of points outside the envelope (%) is less than five for model 1, considering both distances. This fact did not occur for the null model. Also, the distribution of these values within each

level appears to be symmetric and has approximately the same variability (small one) (Figures 3.4(a) and 3.4(b)). Therefore, most of the time, model 1 was adequate to describe the data based on a half-normal plot compared to the null model in this scenario. The results of the descriptive statistics for the null model and model 1 based on the 1000 simulations are displayed in Table 3.2

Table 3.2. Descriptive statistics referring to the number (nº) of points outside the simulated envelope for the null and 1 models, fitted to the simulated grouped data, considering the Euclidean and Mahalanobis distances.

Distance	Model	Descriptive	Nº of points outside
Euclidean	Null	Mean	99.314
		Standard deviation	1.286
		Outliers	12
	1	Mean	2.992
		Standard deviation	5.423
		Outliers	82
Mahalanobis	Null	Mean	35.480
		Standard deviation	8.095
		Outliers	3
	1	Mean	4.014
		Standard deviation	6.006
		Outliers	73

The results in Table 3.2 confirm that model 1 describes well the observed data in this simulation study. The outliers for model 1 in both distances reflect a vast departure from the behavior of most of the points. That is, the points outside the envelope far above 5%. For the null model, the outliers did not change the fact that the number of points outside was more than 5% in practically all simulations (except for one point). Then, the null model was unsuitable for the data. It is also observed that the Euclidean distance presented the highest number of outliers.

Similar conclusions can be observed for the scenario under model 2 for grouped data with 1000 simulations, given that the median of points outside the envelope (%) is less than five for model 2 and more than five for the null model at both distances, Figures 3.5(a) and 3.5(b).

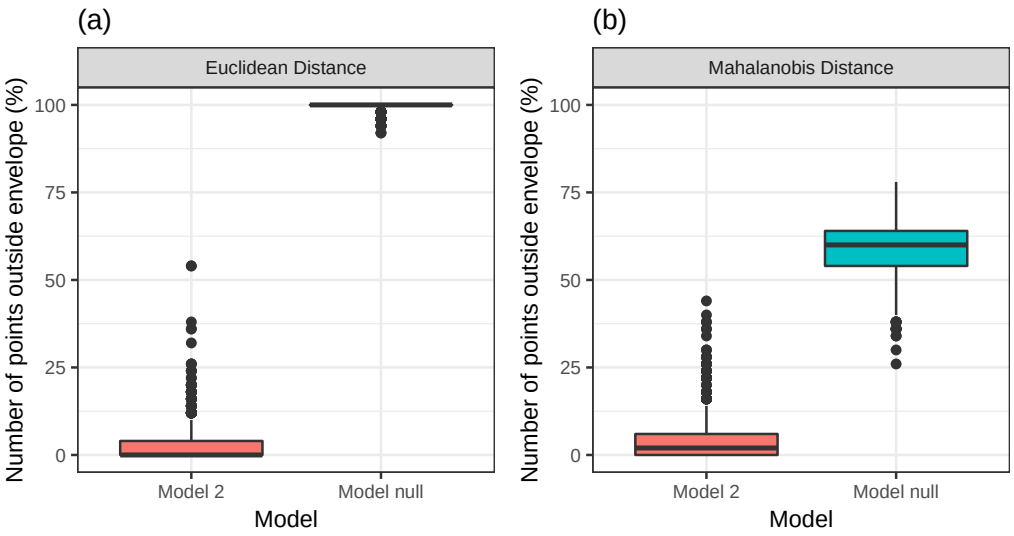

Figure 3.5. Boxplots of the number of points outside the simulated envelope of model 2 (continuous and factor covariates) and null model (intercept effect) referring to (a) Euclidean and (b) Mahalanobis distances.

The results of descriptive statistics for model 2 and the null model, Table 3.3, confirm the obtained conclusions using Euclidean and Mahalanobis distances. In addition, the Euclidean distance again presents more number of outliers.

Table 3.3. Descriptive statistics referring to the number (nº) of points outside the simulated envelope for the null model and model 2, fitted to the simulated grouped data, considering the Euclidean and Mahalanobis distances.

Distance	Model	Descriptive	Nº of points outside
Euclidean	Null	Mean	99.456
		Standard deviation	1.167
		Outliers	215
	2	Mean	2.824
		Standard deviation	5.663
		Outliers	82
Mahalanobis	Null	Mean	58.484
		Standard deviation	7.422
		Outliers	16
	2	Mean	3.766
		Standard deviation	6.077
		Outliers	54

When analyzing the simulation results of the scenarios, it is verified that the standardized randomized quantile residuals demonstrated to follow approximately a standard normal distribution under the correct specification of the model by the Shapiro-Wilk normality test. The proposed methodology for reducing the dimension of the ordinary residuals vector by Euclidean and Mahalanobis distance measures was also satisfactory, identifying the correct model through the half-normal plot with a simulated envelope. These results indicated that standardized randomized quantile residuals (for individual data) and distances (for grouped data) work reasonably well to detect the correct specification of linear predictor structure in the model: continuous covariate or continuous and factor covariates, contributing to the validation step in this study.

3.7 Applications

In this section, two studies of motivation available in the literature are considered to illustrate the procedures presented in sections 3.5 and 3.6.

3.7.1 Application Study 1 - Wine Classification

This first illustration of the procedures refers to a data set from a study carried out by Forina et al. (1988), which became known in the literature involving classification techniques (Jing et al., 2010, Ahammed and Abedin, 2018, among others). In this study, a chemical analysis was carried out at the Institute of Pharmaceutical and Food Analysis and Technologies about 178 wines from three cultivars from the Liguria region in Italy, whose objective was to classify the different cultivars. The response variable represents the type of cultivar, assuming values in the set $\{1, 2, 3\}$. In the analysis, the amounts of 13 chemical constituents of each cultivar were determined, among which are magnesium and phenols that can be considered good indicators of the wine origin (Kallithraka et al., 2001). Further details as well as the dataset are available in the package *ralltle.data* (Graham, 2011) of the *software* R (R Core Team, 2020).

First, an exploratory analysis was carried out to know the behavior of the data. The wine counts in each of the response categories can be seen in Figure 3.6. The majority of wines belong to

cultivar 2 with 71 wines, while the number of wines that belong to cultivar 1 is 59, and cultivar 3 is equal to 48.

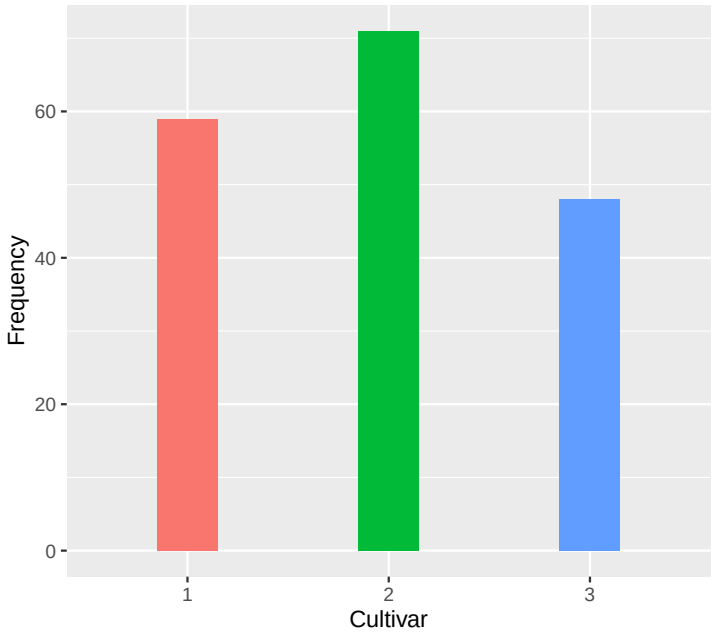

Figure 3.6. Frequencies of wines derived from cultivars 1, 2 and 3 in the study carried out by Forina et al. (1988)

In the selection step of the linear predictor, the covariates phenols and magnesium are considered in the generalized logit model, with the sequential models tested through the likelihood ratio test (LRT), Table 3.4.

Table 3.4. Selection result of the linear predictor between the sequential models through the likelihood ratio test for wine data in the study of Forina et al. (1988)

Model	Linear predictor	Comparation	LRT	$\neq$ *df.	p-value
1 (Null)	Intercept	-	-	-	-
2	Phenols	Model 1 $\times$ Model 2	123.983	2	< 0.01
3	Magnesium + Phenols	Model 2 $\times$ Model 3	13.146	2	< 0.01
4	Magnesium*Phenols	Model 3 $\times$ Model 4	1.251	2	0.535

*df. - degrees of freedom

According to the results of Table 3.4, model 3 is selected to describe the wine data at a significance level of 5%, so the wine classification depends on the phenols and magnesium covariates. The Akaike Information Criterion (AIC) can be used to compare models, where the lowest value indicates a more parsimonious fit. The AIC values for models 1, 2, 3 and 4 are, respectively, 390.63; 270.65; 261.50 and 264.25. The lowest AIC value was for model 3, but this measure does not verify the goodness-of-fit of the model or validate the assumption of response variable distribution.

The estimates, standard errors, as well as 95 % confidence intervals (CI) of the model parameters with the effects of phenols and magnesium, are presented in Table 3.5.

Table 3.5. Parameters estimates of generalized logit model with phenols and magnesium effects selected in the wine data analysis in the study of Forina et al. (1988).

Parameters	Estimate	Standard error	CI	
			2.5%	97.5%
Intercept 2	11.986	2.160	7.752	16.219
Intercept 3	14.180	2.674	8.938	19.422
Phenols 2	-2.438	0.527	-3.471	-1.407
Magnesium 2	-0.056	0.017	-0.089	-0.021
Phenols 3	-5.486	0.762	-6.978	-3.993
Magnesium 3	-0.021	0.020	-0.061	0.018

The estimated classifications of wine by model 3 were obtained for each of the cultivars and calculated their proportions. It is possible to verify the observed and estimated proportions of wines for each cultivar in Figure 3.7, in which there are visual indications that the estimated values are close to the observed ones, evidencing that the model is well fitted.

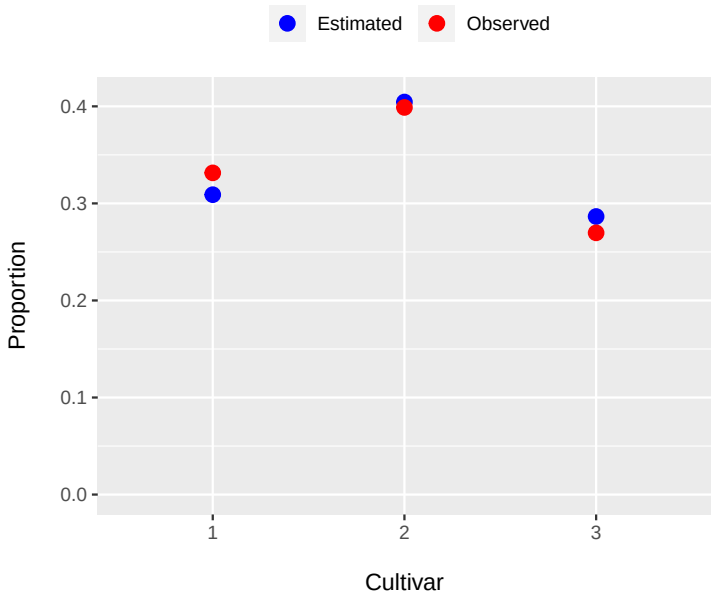

Figure 3.7. Observed and estimated proportions of wines in each cultivar by model 3 in the study of Forina et al. (1988).

In the residual analysis, the histogram of the standardized randomized quantile residuals, Figure 3.8, visually has a shape similar to the standard normal distribution. The p-value of the Shapiro-Wilk test is approximately 0.167, which indicates in favor of the hypothesis that the standardized randomized quantile residuals follow a standard normal distribution. These residuals presented mean 0 and variance 1 because of standardization, while skewness and kurtosis were equal to 0.28 and 2.76, respectively. These values are relatively close to those expected for the normal distribution, which has skewness equal to 0 and kurtosis with a value equal to 3.

The half-normal plot with a simulated envelope for the standardized randomized quantile residuals is showing at (Figure 3.9 (a)), it can be concluded that there is evidence that the model fits the data well since no point is outside the envelope. The scatterplot of residuals versus fitted values can be observed in Figure 3.9 (b), in which residuals vary mainly between -2 and 2 and no pattern is evident, which suggests a good fit of model 3 (phenols and magnesium effects) to the data.

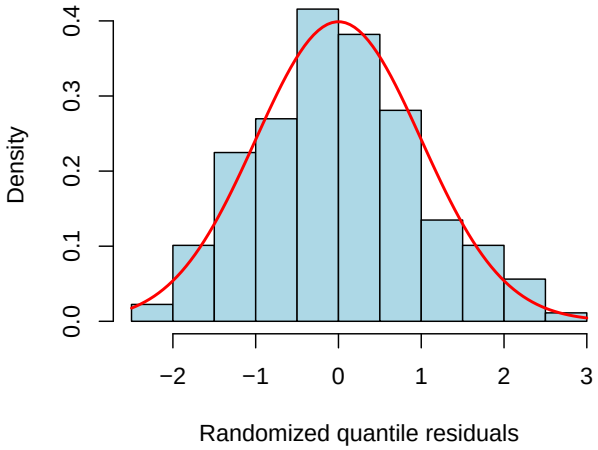

Figure 3.8. Histogram of standardized randomized quantile residuals to assess the goodness-of-fit of generalized logit model with phenols and magnesium effects selected in the wine data analysis in the study of Forina et al. (1988).

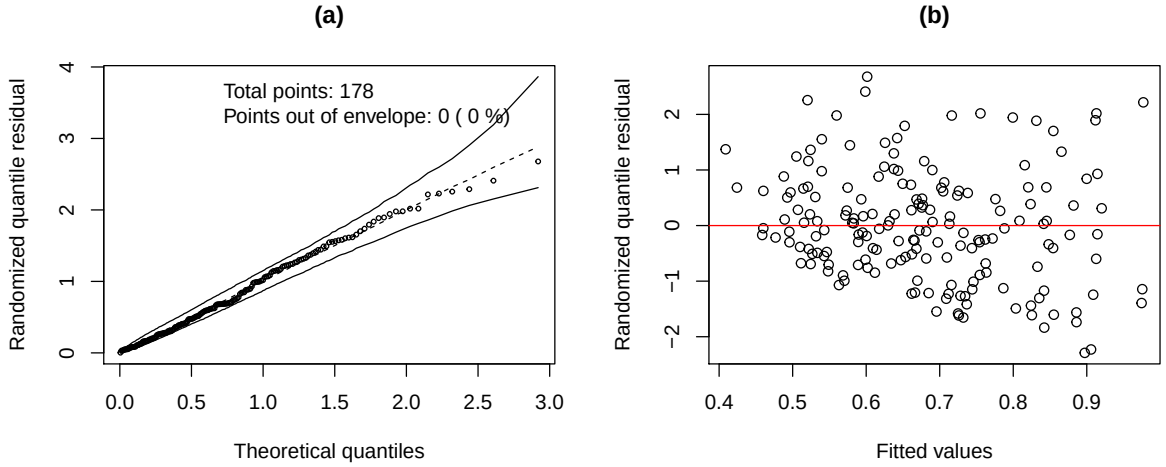

Figure 3.9. a) Half-normal plot with a simulated envelope (confidence level = 95%) of the standardized randomized quantile residuals and b) Residuals versus fitted values plot to model 3 in the wine data analysis in the study of Forina et al. (1988).

3.7.2 Application Study 2 - Preference for the student program of high school students

The data set for this application refers to the choice made by high school students among different programs. This sample of 200 individuals was available in 2013 by the statistical consulting group at the University of California at Los Angeles (UCLA), being used by Molina et al. (2015) and Abonazel and Farghali (2018) in studies involving the estimation of model parameters for polytomous data. Furthermore, it is widely used as an example in statistical software packages for multinomial regression (Dalzell and Reiter, 2018). The response variable is the choice by a program, assuming values in the set $\{1 - \text{academic}, 2 - \text{general}, 3 - \text{vocational}\}$. Among the 11 covariates available in this study are socioeconomic status, gender, and scores in specific subjects (mathematics, social studies, and writing,

among others). This data set can be obtained by accessing the internet site in UCLA (2021). For the present application, the covariate considered is the math score of the student, corresponding to the continuous covariate. Then, the data were organized in a grouped structure with a sample size equal to 40, Table 3.6, for which it was verified whether the mathematics scores contributed to the decision of the new students between the different programs.

Table 3.6. Contingency table obtained from data set of UCLA (2021) - frequencies of program choices (Academic, General, and Vocational) using math scores of high school students

Score Math	Student Program			Total
	Academic	General	Vocation	
33	0	0	1	1
35	1	0	0	1
37	0	0	1	1
38	0	1	1	2
39	2	0	4	6
40	0	0	10	10
41	2	2	3	7
42	5	1	1	7
43	3	4	0	7
44	1	1	2	4
45	1	4	3	8
46	4	1	3	8
47	0	0	3	3
48	1	4	0	5
49	3	6	1	10
50	1	5	1	7
51	0	5	3	8
52	1	2	3	6
53	0	5	2	7
54	3	6	1	10
55	2	2	1	5
56	3	2	2	7
57	4	7	2	13
58	3	3	0	6
59	0	2	0	2
60	2	3	0	5
61	2	5	0	7
62	0	4	0	4
63	1	4	0	5
64	0	5	0	5
65	0	3	0	3
66	0	3	1	4
67	0	2	0	2
68	0	1	0	1

Continued on next page

69	0	2	0	2
70	0	1	0	1
71	0	4	0	4
72	0	3	0	3
73	0	1	0	1
75	0	1	1	2
Total	45	105	50	200

Through the exploratory analysis, it was possible to verify the behavior of this data set, which presents variability in the choices of programs using the math scores of students, as shown in Figure 3.10. The general program presents a major occurrence, being practically predominant among the new students who obtained scores higher than 58, while the students with scores lower than 42 (except for a score of 35) and a score equal to 47 present more frequency in the vocational program. Finally, higher frequencies concerning the academic program refer to scores equal to 35, 42, 46, and 56.

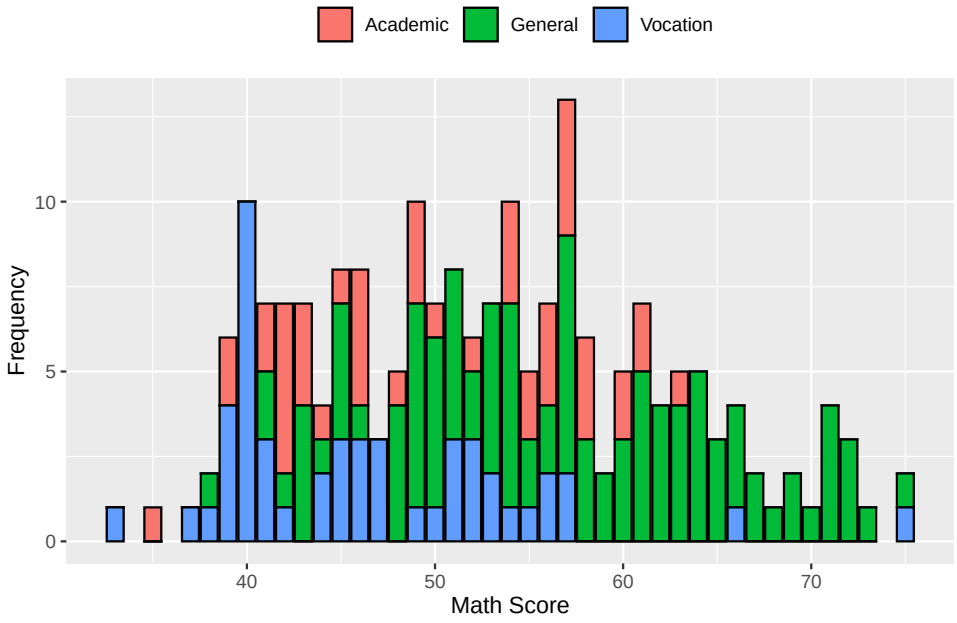

Figure 3.10. Histogram of relative frequencies to academic, general, and vocational program choices by math grades in high school students data obtained from UCLA (2021).

Considering the null hypothesis that the program choice made by high school students is independent of the score in mathematics, the likelihood ratio test compares the null model (intercept effect) with model 1, which considers the effect of the continuous covariate. Model 1 was selected to describe the data at a 5% significance level ($\text{LRT} = 51,965$ and p-value < 0.01). Model 1 also presented a lower AIC (182.81) than the null model (230.77), which shows a better fit. Based on this result, it is concluded that the score in mathematics is significant in the program choice made by the new students in the data available by UCLA (2021).

The estimates, standard errors, as well as 95 % confidence intervals (CI) of model 1 with the math score effect are presented in Table 3.7.

Table 3.7. Parameter estimates of model 1 with the effects of the score in mathematics selected for the analysis of data from high school students available in UCLA (2021).

Parameter	Estimate	Standard error	CI	
			2.5%	97.5%
Intercept 2	-4.055	1.218	-6.443	-1.668
Intercept 3	3.136	1.362	0.466	5.806
Math score 2	0.092	0.023	0.047	0.137
Math score 3	-0.062	0.028	-0.118	-0.008

Thus, the expressions in logits terms for the model 1 are given by

$$\log\left[\frac{\pi_2(\mathbf{x})}{\pi_1(\mathbf{x})}\right] = -4.055 + 0.0928\,x \quad \text{and} \quad \log\left[\frac{\pi_3(\mathbf{x})}{\pi_1(\mathbf{x})}\right] = 3.136 - 0.062\,x.$$

em where x represents the continuous covariate (math score).

It was used the sorted absolute values of the diagnostic measures given by the Euclidean and Mahalanobis distances to detect the presence of outliers. The behavior of these values in the half-normal plot with a simulated envelope is presented in Figures 3.11 and 3.12.

The values of Euclidean distance (Figure 3.11) for model 1 are mainly inside the simulated envelope (one point outside), indicating that this model satisfactorily fitted the data.

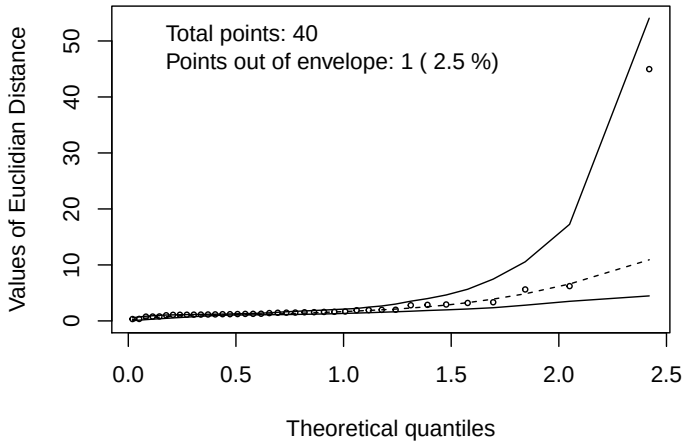

Figure 3.11. Half-normal plot with a simulated envelope (confidence level = 95%) of Euclidean distance points for model with math score effect for the data available in UCLA (2021).

In the same way, the half-normal plot of the Mahalanobis distance points (Figure 3.12) of the math score-effect model showed evidence that this model is adequate for analyzing the data, with no points outside the simulated envelope.

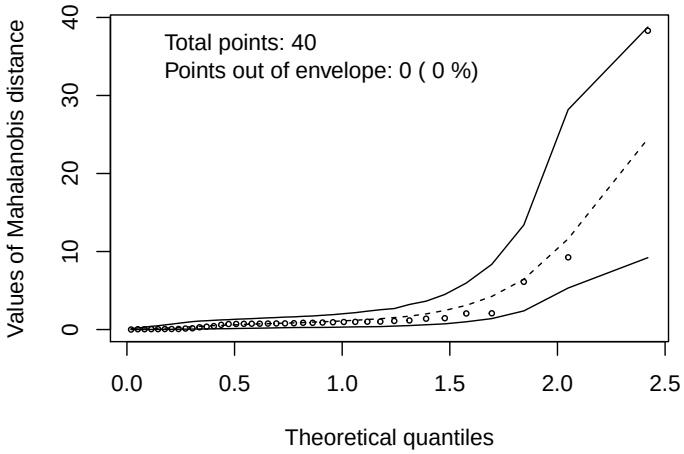

Figure 3.12. Half-normal plot with a simulated envelope (confidence level = 95%) of Mahalanobis distance points for model with math score effect for the data available in UCLA (2021).

3.8 Conclusion

In this chapter, the standardized randomized quantile residuals and the reduction of the ordinary residuals using distance measures were obtained to perform the diagnostics of the generalized logit model for nominal data with individual and grouped structures. These residuals and the proposed distances presented good performance in assessing the goodness-of-fit of the model with continuous covariate or continuous and factor covariates, particular situations that can occur in many areas of knowledge. This approach allows evaluating the simultaneous fit of the model in the different data structures with a moderate sample size. Studies under a small sample size are necessary to assess the fit of the model, which could lead to sampling uncertainty in the residuals and a singular matrix for Mahalanobis distance. Also, future simulations studies can be done to check the normality of randomized quantile residuals for the grouped data structure. Based on the applications, the obtained results revealed non-violation in model assumptions or outliers by the half-normal plot with a simulated envelope, the plot of residuals versus fitted values, and the Shapiro-Wilk test. In this way, the randomized quantile residuals and the distances can be potential alternatives to evaluate the fit of the generalized logit models.

References

Abonazel, M. R. and Farghali, R. A. (2018). Liu-type multinomial logistic estimator. *Sankhya B: The Indian Journal of Statistics*, 81(2):203–225.

Agresti, A. (2002). *An introduction to categorical data analysis*. John Wiley & Sons, Nova Jersey, 3 edition.

Agresti, A. (2007). *An introduction to categorical data analysis*. John Wiley & Sons, Hoboken, New Jersey, 2 edition.

Ahammed, B. and Abedin, M. (2018). Predicting wine types with different classification techniques. *Model Assisted Statistics and Applications*, 13(1):85–93.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Atkinson, A. C. (1985). Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis. Technical report.

Bilder, C. R. and Loughin, T. M. (2014). *Analysis of categorical data with R*. Chapman and Hall/CRC Press, Boca Raton, 1 edition.

Cheng, C., Wang, R., and Zhang, H. (2021). Surrogate residuals for discrete choice models. *Journal of Computational and Graphical Statistics*, 30(1):67–77.

Cook, R. D. and Tsai, C. L. (1985). Residuals in nonlinear regression. *Biometrika*, 72(1):23–29.

Dalzell, N. M. and Reiter, J. P. (2018). Regression modeling and file matching using possibly erroneous matching variables. *Journal of Computational and Graphical Statistics*, 27(4):728–738.

Davis, A. (2018). Package 'pmultinom'. *R package version*.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.

Feng, C., Li, L., and Sadeghpour, A. (2020). A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, 20(1):1–21.

Forina, M., Lanteri, S., and Armanino, C. (1988). Parvus - an extendible package for data exploration, classification and correlation. *Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy*.

Ghorbani, H. (2019). Mahalanobis distance and its application for detecting multivariate outliers. *Facta Univ Ser Math Inform*, 34(3):583–95.

Graham, J. W. (2011). *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. Use R! Springer.

Gupta, A. K., Nguyen, T., and Pardo, L. (2008). Residuals for polytomous logistic regression models based on φ-divergences test statistics. *Statistics*, 42(6):495–514.

Hadi, A. S., Rahmatullah Imon, A. H. M., and Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):57–70.

Hossain, M. and Islam, M. A. (2003). Application of local influence diagnostics to the linear logistic regression models. *Dhaka University Journal of Science*, 51(2):269–278.

Jing, L., Jin-Jia, W., Tao, Z., Chong-Xiao, M., and Wen-Xue, H. (2010). The graphical feature extraction of star plot for wine quality classification. In *2010 First International Conference on Pervasive Computing, Signal Processing and Applications*, pages 771–774. IEEE.

Johnson, R. A. and Wichern, D. W. (2007). *Applied multivariate statistical analysis*, volume 6. Pearson Prentice Hall.

Kallithraka, S., Arvanitoyannis, I., Kefalas, P., El-Zajouli, A., Soufleros, E., and Psarra, E. (2001). Instrumental and sensory analysis of greek wines; implementation of principal component analysis (pca) for classification according to geographical origin. *Food Chemistry*, 73(4):501–514.

Kannan, K. S. and Manoj, K. (2015). Outlier detection in multivariate data. *Applied Mathematical Sciences*, 47(9):2317–2324.

Klar, B. and Meintanis, S. G. (2012). Specification tests for the response distribution in generalized linear models. *Computational Statistics*, 27(2):251–267.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79(385):61–71.

Levin, B. (1981). A representation for multinomial cumulative distribution functions. *The Annals of Statistics*, pages 1123–1126.

Liu, D. and Zhang, H. (2018). Residuals and diagnostics for ordinal regression models: A surrogate approach. *Journal of the American Statistical Association*, 113(522):845–854.

Maesschalck, R. d., Jouan-Rimbaud, D., and Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18.

Molina, D., Rueda, M. M., Arcos, A., and Ranalli, M. G. (2015). Multinomial logistic estimation in dual frame surveys. *SORT*, 39(2):309–336.

Moral, R. A., Hinde, J., and Demétrio, C. G. B. (2017). Half-normal plots and overdispersed models in r: the hnp package. *Journal of Statistical Software*, 81(1):1–23.

Nobre, J. S. and Singer, J. M. (2007). Residual analysis for linear mixed models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 49(6):863–875.

Pereira, G. H. A. (2019). On quantile residuals in beta regression. *Communications in Statistics-Simulation and Computation*, 48(1):302–316.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of statistics*, 9(4):705–724.

R Core Team, . (2020). R: A language and environment for statistical computing.

Reiter, J. P. and Kohnen, C. N. (2005). Categorical data regression diagnostics for remote access servers. *Journal of Statistical Computation and Simulation*, 75(11):889–903.

Ripley, B. and Venables, W. (2016). Package 'nnet'. *R package version*, 7(3-12):700.

Royston, P. (1995). Remark as r94: A remark on algorithm as 181: The w-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):547–551.

Seber, G. and Nyangoma, S. (2000). Residuals for multinomial models. *Biometrika*, 87(1):183–191.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Sharma, S. (1996). *Applied multivariate techniques*. John Wiley & Sons.

Silva, J. A. P. (2003). *Métodos de diagnóstico em modelos logísticos trinomiais*. Dissertação (mestrado em estatística).

Tang, W., Hua, H., and Tu, X. M. (2012). *Applied categorical and count data analysis*. Chapman and Hall/CRC Press, Boca Raton, 1 edition.

Tutz, G. (2011). *Regression for categorical data*, volume 34. Cambridge University Press, Cambridge.

UCLA, S. C. G. (2021). Hsbdemo data set. `https://stats.idre.ucla.edu/stat/data/hsbdemo.dta`.

Zelterman, D. (2015). *Applied multivariate statistics with R*. Springer.

Appendix

```
#############################################################################
#R code
#############################################################################
#This script is for the simulations with continuous covariate in the model
(analogue way to continuous and factor covariates)
#############################################################################################
rm(list=ls(all=TRUE))
require(nnet) #Fit generalized logit Model
require(hnp) #half-normal plot using envelope simulation
require(pmultinom) #Calculate cdf of multinomial dist
require(stats) #To obtain Mahalanobis and Euclidean Distance
require(ggplot2);require(dplyr);require(tidyr);require(moments)
#############################################################################################
#RQRs for multinomial case:
#RQR for correct model
RQR.r <- function(m,y,pred.r){
  n<-nrow(y)
  res.quantile.r<-vector()
  for (i in 1:n) {
    if(y[i,1]==1 && y[i,2]==0 && y[i,3]==0){
      res.quantile.r[i] <- qnorm(pmultinom(upper=c(y[i,1]-1,m,m),size = m,
      probs=pred.r[i,],method="exact")+
      dmultinom(y[i,],size = m, pred.r[i,])*runif(1))
       }else if(y[i,1]==0 && y[i,2]==1 && y[i,3]==0){
      res.quantile.r[i] <- qnorm(pmultinom(upper=c(m,y[i,2]-1,m),size = m,
      probs=pred.r[i,],method="exact")+
      dmultinom(y[i,],size = m, pred.r[i,]) * runif(1))
```

```
    }else (y[i,1]==0 && y[i,2]==0 && y[i,3]==1){
      res.quantile.r[i] <- qnorm(pmultinom(upper=c(m,m,y[i,3]-1),size = m,
      probs=pred.r[i,],method="exact")+
      dmultinom(y[i,],size = m, pred.r[i,]) * runif(1))
   }}
  return(res.quantile.r)
  }

#RQR for null model
RQR.w <- function(m,y,pred.w){
  n<-nrow(y)
  res.quantile.w<-vector()
  for (i in 1:n) {
    if(y[i,1]==1 && y[i,2]==0 && y[i,3]==0){
      res.quantile.w[i] <- qnorm(pmultinom(upper=c(y[i,1]-1,m,m),size = m,
      probs=pred.w[i,],method="exact")+
     dmultinom(y[i,],size = m, pred.w[i,]) * runif(1))
     }else if(y[i,1]==0 && y[i,2]==1 && y[i,3]==0){
      res.quantile.w[i] <- qnorm(pmultinom(upper=c(m,y[i,2]-1,m),size = m,
      probs=pred.w[i,],method="exact")+
      dmultinom(y[i,],size = m, pred.w[i,]) * runif(1))
      }else(y[i,1]==0 && y[i,2]==0 && y[i,3]==1){
      res.quantile.w[i] <- qnorm(pmultinom(upper=c(m,m,y[i,3]-1),size = m,
      probs=pred.w[i,],method="exact")+
      dmultinom(y[i,],size = m, pred.w[i,]) * runif(1))
   }}
  return(res.quantile.w)
}
##############################################################################
#Function to the simulations
f <- function(m,x,prob) {
  y <- t(apply(prob, 1, rmultinom,n=1, size = m))
   dfM <- data.frame(x,y)
   #Fit of model
  fit.w <- multinom(y~ 1,data = dfM,trace="FALSE") #null model
  fit.r <- multinom(y ~ x,data = dfM,trace="FALSE") #correct model
   p_value <- anova(fit.w, fit.r)[2,7] #p-value of the test
  #Predicted probabilities
  pred.w<-predict(fit.w, type = "prob")
  pred.r<-predict(fit.r, type = "prob")
  # Standardized Residuals
  res.w<-RQR.w(m,y,pred.w)
  res.r<-RQR.r(m,y,pred.r)
  res.mult.w.n<- (res.w-mean(res.w))/sd(res.w)
  res.mult.r.n<-(res.r-mean(res.r))/sd(res.r)
 #Half-normal plot with simulated envelope
invisible(capture.output(myhnp.w<-hnp(res.mult.w.n, print = T,scale=T,
```

```
plot.sim = "FALSE",sim = 1000)))
invisible(capture.output(myhnp.r<-hnp(res.mult.r.n, print=T,scale=T,
plot.sim = "FALSE",sim = 1000)))
#Percentage of points outside the envelope
 npoints.w<-myhnp.w$out
  perc.w<-round((npoints.w/myhnp.w$total)*100,2)
  npoints.r<-myhnp.r$out
  perc.r<-round((npoints.r/myhnp.r$total)*100,2)
  ### Shapiro Wilk Test
  test.w<-shapiro.test(res.mult.w.n)$p.value
  test.r<-shapiro.test(res.mult.r.n)$p.value
  #Descriptive statistics
  #null model
  mean.w <- mean(res.mult.w.n)
  sd.w<- sd(res.mult.w.n)
  kurt.w<- kurtosis(res.mult.w.n)
  skew.w<-skewness(res.mult.w.n)
 #correct model
  mean.r <- mean(res.mult.r.n)
  sd.r<- sd(res.mult.r.n)
  kurt.r<- kurtosis(res.mult.r.n)
  skew.r<-skewness(res.mult.r.n)
  #To show the results
   Values_Model<-c("Perc.w"=perc.w,"Perc.r"=perc.r)
  Value_test<-c("SW_test.w"=test.w,"SW_test.r"=test.r,"p-value"=p_value)
  Stats.w <-c("media.w"=mean.w,"sd.w"=sd.w,"kurt.w"=kurt.w,"skew.w"=skew.w)
  Stats.r<-c("media.r"=mean.r,"sd.r"=sd.r,"kurt.r"=kurt.r,"skew.r"=skew.r)
   Result_final<-c(Values_Model,Value_test,Stats.w,Stats.r)
   return(Result_final)
 }

#Residual using multinomial
#3 categories and continuous covariate
n<-100# sample size
x<-rnorm(n) #covariate
#first category as reference
z2<--1.38-2.7*x
z3<- 3.51-5.11*x
#to obtain the probabilities
den<-1+exp(z2)+exp(z3)
p1<-1/den; p2<-exp(z2)/den; p3<-exp(z3)/den
prob<-cbind(p1,p2,p3)
#number of simulation
n_replic<-1000
m<-1
sim_scenario <- replicate(n_replic,f(m,x,prob))
##############################################################################
```

```
#Euclidean Mahalanobis distances for multinomial case:
#Function to the simulations
f <- function(m,x,prob) {
  y <- t(apply(prob, 1, rmultinom,n=1, size = m))
  # Value of Y and X together
  dfM <- cbind.data.frame(y, x)
  #Fit of model
  fit.w <- multinom(y ~ 1, data = dfM,trace=FALSE) #null model
  fit.r <- multinom(y ~ x, data = dfM,trace=FALSE) #correct model
  p_value <- anova(fit.w, fit.r)[2,7] #p-value of the test
  #hnp for redution of ordinary residuals by the distances
  ##Euclidean distance
  d_eucl<- function(obj) {
    r <- resid(obj)
    l2_r <- apply(r, 1, function(x) dist(rbind(x,rep(0,length(x)))))
    return(as.numeric(l2_r))
  }
  #Mahalanobis distance
  d_Mahal <- function(obj) {
    r <- resid(obj)
    k<-ncol(r)
    Sr<-solve(cov(r))
    D2 <- mahalanobis(r, rep(0,nrow(r)),Sr,inverted = T)
    return(D2)
  }
  #Implementing new class for hnp
  n<-length(x)
  s_fun <- function(n, obj) {
    pred<-predict(obj, type = "prob")
    newresp<- t(apply(pred, 1, function(x) rmultinom(1, size = m, x)))
    newresp
  }
  f_fun.w <- function(newresp) {
    multinom(newresp ~ 1, data = dfM)
  }
  f_fun.r <- function(newresp) {
    multinom(newresp ~ x, data = dfM)
  }
  #hnp for null model vs correct model
  #Euclidian Dist
  #Use of this function for suppressing convergence message
  invisible(capture.output(my_hnpE.w<-hnp(fit.w, newclass = TRUE,
  diagfun = d_eucl,simfun = s_fun, fitfun = f_fun.w, how.many.out = T,
  plot.sim = "FALSE")))
  invisible(capture.output(my_hnpE.r<-hnp(fit.r, newclass = TRUE,
  diagfun = d_eucl,simfun = s_fun, fitfun = f_fun.r, how.many.out = T,
  plot.sim = "FALSE")))
```

```
  #Percentage of points outside the envelope
  n_pointsE.w<-my_hnpE.w$out
  percE.w<-round((n_pointsE.w/my_hnpE.w$total)*100,2)
    n_pointsE.r<-my_hnpE.r$out
  percE.r<-round((n_pointsE.r/my_hnpE.r$total)*100,2)
  #Mahalanobis Dist
  invisible(capture.output(my_hnpM.w<-hnp(fit.w, newclass = TRUE,
  diagfun = d_Mahal, simfun = s_fun, fitfun = f_fun.w, how.many.out = T,
  plot.sim = "FALSE")))

  invisible(capture.output(my_hnpM.r<-hnp(fit.r, newclass = TRUE,
  diagfun = d_Mahal, simfun = s_fun, fitfun = f_fun.r, how.many.out = T,
  plot.sim = "FALSE")))
   #Percentage of points outside the envelope
  n_pointsM.w<-my_hnpM.w$out
  percM.w<-round((n_pointsM.w/my_hnpM.w$total)*100,2)
  n_pointsM.r<-my_hnpM.r$out
  percM.r<-round((n_pointsM.r/my_hnpM.r$total)*100,2)
  #To show the results
  Incorrect_Cor_Model<-c("Eucl_Inc"=percE.w,"Eucl_Cor"=percE.r,
  "Mahal_Inc"=percM.w, "Mahal_Cor"=percM.r)
  Value_test<-c("p-value"=p_value,"LR"=lr_stat )
   Result_final<-c(Incorrect_Cor_Model, Value_test)
   return(Result_final)
}
#3 categories and continuous covariate
n<-50#number of samples
x<-rnorm(n) #covariate
#Assigning the values of intercepts and betas
#The first level is assigned the role of reference.
z2<--1.38-2.7*x
z3<- 3.51-5.11*x
den<-1+exp(z2)+exp(z3)
p1<-1/den; p2<-exp(z2)/den; p3<-exp(z3)/den
prob<-cbind(p1,p2,p3)
m<-5 #After 10 and 15
n_replic<-1000 #number of simulations
sim_scenario <- replicate(n_replic,f(m,x,prob))
##############################################################################
```

4 FINAL CONSIDERATIONS

This thesis is an introduction to residual analysis and diagnostics for categorized data. A review of the available residuals in the literature was carried out in Chapter 2, emphasizing the surrogate residuals and presenting an application for the case in which the categorized variable is ordinal with an individual structure. However, this residual cannot be applied to cases with nominal responses or a grouped structure for the data.

The specific contributions of this thesis are in Chapter 3. The first is the normality investigation of the randomized quantile residuals associated with individual nominal data through simulation studies. The second contribution is the proposal of Euclidean and Mahalanobis distances in the dimension reduction of the vector of ordinary residuals in the nominal grouped case. These measures were used to detect outliers in the diagnostics of multinomial regression. Under simulation studies in different scenarios, the approaches showed promising results in assessing the goodness-of-fit of the generalized logit model. Also, the applications presented reinforce this conclusion. The possible limitations are related to small samples. The statistical power of the randomized quantile residual may be low in small samples to detect the incorrect specification of the model. For the cases using the Mahalanobis distance, small samples can make difficult the calculation of the covariance matrix. Despite that, both approaches have substantive appeal in the diagnostics of generalized logit models for nominal data (individual or grouped) with moderate sample sizes, reasonably detecting the misspecification linear predictor in this study: continuous covariate or continuous and factor covariates.

For future research, new scenarios can be studied to detect many forms of model misspecification such as non-linearity or interaction effect under distinct sample sizes as well as realize simulations studies for normality of randomized quantile residuals in the grouped data structure to evaluate the fit of the generalized logit models. In addition, a sensitivity analysis can be performed to assess the goodness-of-fit of these models and explore the diagnostics in Bayesian inference.