Brazilian Journal of Biometrics, **41**, 287–310 (2023)
doi:10.28951/bjb.v41i3.637

BRAZILIAN JOURNAL OF
BIOMΣTRICS
ISSN:2764-5290

ARTICLE

Ordinal data and residual analysis: Review and application

Patricia Peres Araripe* and Idemauro Antonio Rodrigues de Lara

Department of Exact Sciences, "Luiz de Queiroz" College of Agriculture, University of São Paulo, Piracicaba-SP, Brazil
*Corresponding author. Email: araripe.patricia@gmail.com

(Received: November 1, 2022; Revised: January 3, 2023; Accepted: February 3, 2023; Published: September 25, 2023)

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 subject. 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.

1. Introduction

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, 2010). When the proportionality assumption is not valid, the cumulative logit model (Williams & Grizzle, 1972) can be fitted to the data or the adjacent-categories logit model, for example (Ananth & 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 & Weisberg, 1982). McCullagh & 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 subject (Reiter & 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 subject 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 & Shepherd (2012) and Liu & Zhang (2018) defined residuals that correspond to a single value per subject regardless of the number of categories. While the residual proposed by Li & Shepherd (2012) is obtained in the discrete space of the original response, in the approach used by Liu & Zhang (2018), a continuous variable replaces the ordinal variable, and the residual is defined through this new variable. Liu & Zhang (2018) compared the performance of the residuals so-called surrogate, with those proposed by Li & 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 & 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. Models for ordinal response

The multinomial distribution is the most important and usual one for random variables associated with categorical data. Thus, it is the distribution that is assumed, except for overdispersion, in classic models with polytomous responses (Agresti, 2002). 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. The components of the random vector $\mathbf{Y} = (Y_1, \ldots, Y_J)'$ give the cell counts in categories $1, \ldots, J$. Then, the random vector 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

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

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 $y_{j'}$, $\forall j \neq j'$, $j' = 1, \ldots, J$, is obtained by $\mathrm{Cov}(Y_j, Y_{j'}) = -m\pi_j\pi_{j'}$, and that the marginal distribution of each y_j is binomial.

2.1 Cumulative logit model

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

This model is defined by:

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

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is the vector of the p covariates for the i-th subject, $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})'$ represents the vector of regression parameters and α_j is the intercept, with $j = 1, 2, \ldots, J-1$, these ones constitute the systematic part of model. Here for the multinomial response Y_i, $\gamma_{ij}(\mathbf{x}_i)$ is the cumulative probability of the subject 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 (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 subject 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). Additionally, as is a classic in generalized linear models, under conditions of regularity, in the ordinal case, asymptotically $\hat{\boldsymbol{\theta}}$ has a normal distribution, that is, $\hat{\boldsymbol{\theta}} \sim \mathrm{N}(\boldsymbol{\theta}, \Im^{-1})$, where $\Im$ is the Fisher information matrix. This matrix is of fundamental importance in the construction of hypothesis tests and asymptotic confidence intervals for the elements of $\boldsymbol{\theta}$ (via the Wald method), since the square root of the main diagonal elements are the standard errors of the estimators. More details can be found at McCullagh (1980).

An alternative to model (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 & 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 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

$$\text{logit}\left[\gamma_{ij}(\mathbf{x}_i)\right] = \log\left[\frac{\gamma_{ij}(\mathbf{x}_i)}{1 - \gamma_{ij}(\mathbf{x}_i)}\right] = \alpha_j + \sum_{k=1}^{p} \beta_k x_{ik} = \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i, \quad j = 1, \ldots, J-1, \qquad (2)$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is the vector of covariates for the subject 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 for the multinomial response Y_i, $\gamma_{ij}(\mathbf{x}_i)$ is the cumulative probability of subject 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) 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 & Loughin, 2014). According to Agresti (2007), the maximum likelihood procedure can be used to estimate the parameters of the model (2), with a likelihood function for the random sample of dimension n described by

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

where $y_{ij} = 1$ if the response of subject 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 $\boldsymbol{\theta} = \left(\alpha_1, \ldots, \alpha_{J-1}, \boldsymbol{\beta}\right)'$ 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 & 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 (1), the proportionality assumption can be verified through the likelihood ratio test (LRT) with the following hypotheses

$$\begin{cases} H_0 : \boldsymbol{\beta}_j' = \boldsymbol{\beta}', & \forall j = 1, 2, \ldots, J-1 \\ H_1 : \boldsymbol{\beta}_j' \neq \boldsymbol{\beta}', & forsomej. \end{cases}$$

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) and L_{H_1} is the likelihood function under the alternative hypothesis H_1, i.e., referring to model (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 subject. Globally, all model covariates are considered, while subjectly, 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, 2010).

2.3 Partial proportional odds model

The proportional odds assumption is not always achieved in practice. A model proposed by Peterson & 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 subject 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

$$\text{logit}\left[\gamma_{ij}(\mathbf{x}_i, \mathbf{z}_i)\right] = \log\left[\frac{\gamma_{ij}(\mathbf{x}_i, \mathbf{z}_i)}{1-\gamma_{ij}(\mathbf{x}_i, \mathbf{z}_i)}\right] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\rho}_j'\mathbf{z}_i, \quad j = 1, \ldots, J-1, \tag{3}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ and $\boldsymbol{\rho}_j = (\rho_{j1}, \rho_{j2}, \ldots, \rho_{jq})'$ sare the vectors of regression parameters, α_j is the intercept and the last category taken as a reference. Here, the vector $\boldsymbol{\rho}_j$ 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 $\boldsymbol{\rho}_j$ are estimated. Furthermore, $\gamma_{ij}(\mathbf{x}_i, \mathbf{z}_i)$ is the cumulative probability of subject 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 (3) are obtained in an analogous way to those obtained for models (1) and (2), so

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

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, 2010). Considering $y_{ij} = 1$ if the response of subject 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 (3) can be obtained by maximizing the likelihood function (or its logarithm) given by

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

where $\boldsymbol{\theta}=\left(\alpha_1,\ldots,\alpha_{J-1},\boldsymbol{\beta},\boldsymbol{\rho}_1,\ldots,\boldsymbol{\rho}_{J-1}\right)'$ 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 & 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. We reinforce that all models mentioned here are based on the multinomial distribution for the response variable, i.e., a random part of the model. They differ in the structure of the linear predictor, the systematic part of the model. Additionally, it is possible to find this model and others for ordinal data in Ananth & Kleinbaum (1997), Agresti (2002), Agresti (2007), Agresti (2010), Tutz (2011), Bilder & Loughin (2014), Giolo (2017), among others.

3. Residuals for ordinal data

After fitting a model to the data, it is essential to verify whether its assumptions are satisfied and identify subjects 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).

3.1 Ordinary Residual

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

$$\hat{\mathbf{r}}_i=\mathbf{y}_i-\hat{\boldsymbol{\pi}}_i=\left(y_{i1}-\hat{\pi}_{i1},y_{i2}-\hat{\pi}_{i2},\ldots,y_{iJ}-\hat{\pi}_{iJ}\right)', \qquad (4)$$

where $\mathbf{y}_i=(y_{i1},y_{i2},\ldots,y_{iJ})'$ is the observed vector with $y_{ij}=1$ if the response of the subject 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 subject. This vector may not be informative in the diagnostic techniques for analyzing residuals since its asymptotic distribution is unknown.

3.2 Cumulative Residual

Specifically for the proportional odds model, defined in section 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 subject is expressed by

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

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 subject. 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 & Lin (2005) for the logistic regression model with binary responses. However, diagnostic plots associated with residuals are difficult to interpret.

3.3 LS Residual

Considering the models that assume the assumption of proportionality for the regression parameters, Li & Shepherd (2012) proposed a residual that is a single value per subject, 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 a subject considering the model 2 is obtained by

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

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 & Zhang (2018), the use of this residual is limited to verifying its zero mean under the correct model.

3.4 Surrogate Residual

The residual defined by Liu & Zhang (2018) is also a single value per subject for the models that assumes the proportional odds. Consider the model (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 $u \in$ 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). 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 $\left(\alpha_{j-1}; \alpha_j\right)$ 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

$$R_i^S = S_i - \mathrm{E}_0(S_i|\mathbf{x}_i) = S_i - \mathrm{E}(Z_i|\mathbf{x}_i) = S_i + \boldsymbol{\beta}'\mathbf{x}_i - \int_{-\infty}^{+\infty} u dG(u) \qquad (5)$$

where $\mathrm{E}_0(.)$ and $\mathrm{E}(.)$ denote, respectively, the mean of variables S_i and Z_i. If the model (2) 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 & 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 (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 & Zhang, 2018):

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

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 & 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

$$\begin{cases} H_0 : Q_n(c; R_1, R_2, \ldots, R_n) = G(c), & \forall c \in (-\infty; +\infty) \\ H_1 : Q_n(c; R_1, R_2, \ldots, R_n) \neq G(c), & forsomec. \end{cases}$$

and test statistic

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

where $d_{KS}(Q_n, G) = \sup_{c \in R} |Q_n(c; R_1, R_2, \ldots, R_n) - G(c)|$ 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° 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 behavior interpretation 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 randomly distributed within the envelope and decide whether the observed residuals are consistent with the fitted model. The envelope is simulated with a confidence level of 95% that contains the residuals, i.e., there is evidence of a good fit when the number of points outside the envelope is below or equal to 5%.

In addition, a sensitivity analysis based on a set of tools (such as leverage, case-deletion, or local influence analysis) can be designed to evaluate changes in the fitted model when some perturbation is imposed on the data or assumptions of the model (Singer *et al.,* 2017). This will be briefly presented below given that the focus of this paper is on the residual analysis.

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 subjects with the fit when deleting a particular subject from the sample. Consider $\hat{\theta}$ and $\hat{\theta}_{(i)}$ the estimated maximum likelihood vectors from the sample with all subjects and the sample without subject i, respectively. An indicator of the influence of i-th subject can be calculated by $\hat{\theta} - \hat{\theta}_{(i)}$. If the estimates differ substantially, the subject can be considered influential.

A measure that also can be used to assess the influence of the i-th subject is the likelihood displacement (Cook & Weisberg, 1982). This measure verifies the distance between the two likelihoods, being given by

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

where $l(\hat{\theta})$ and $l(\hat{\theta}_{(i)})$ are, respectively, the likelihood logarithms of the parameters obtained from the sample with all points and the sample without the i-th subject. When it is not possible to obtain an analytical form for LD_i, a quadratic approximation by Taylor series leads to the following result:

$$LD_i = (\hat{\theta} - \hat{\theta}_{(i)})' F(\hat{\theta})(\hat{\theta} - \hat{\theta}_{(i)}),$$

where $F(\theta) = E\left(-\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'}\right)$ is the Fisher information matrix, which is estimated by substituting $\hat{\theta}$. Generally, it is not possible to obtain a closed form for $\hat{\theta}_{(i)}$, and the one-step approximation is used that takes the first iteration of the iterative process by the Fisher score method when it starts at $\hat{\theta}$ (Paula, 2013).

Another indication of an influential observation is through leverage point. The leverage value measures the potential for a subject to have a large effect on the fitted regression line, being defined as a measure of how far a particular case is (based on only predictor values) from the average of all cases. Also, looking at residuals doesn't help to detect leverage points since they don't necessarily fall off the regression surface (Simonoff, 2003).

Finally, there are several techniques that can help in the diagnostics of the model, which are described in several papers such as Junior & Veiga (2020) that assessed the local influence and the likelihood displacement measure for diagnostics in normal and logistic regression models. Details about these diagnostic measures are covered in Cook & Weisberg (1982), McCullagh & Nelder (1989), Turkman & Silva (2000), 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).

5. Matherials and Methods

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 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.

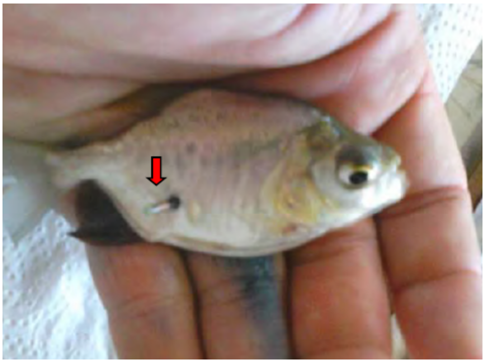

Figure 1. Microchip inserted in the juvenile of Tambaqui in a study carried out by Marques (2018) at the Biofish-Aquicultura farm.

The pituitary gland was collected for gene expression analysis, placed in a stabilizing solution (RNAlater), and stored at -80°C 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, 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.

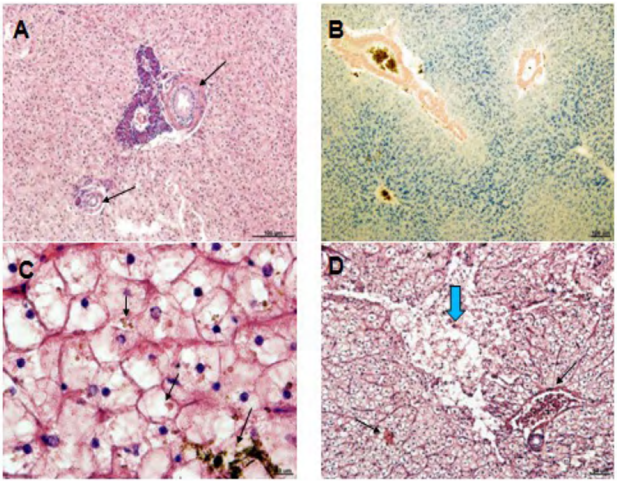

Figure 2. Morphology of the liver tissue of the Tambaqui fish with the histopathological alterations in the experiment carried out by Marques (2018) at the Biofish-Aquicultura farm. A- Ductal hypertrophy (black arrows); B- Hemosiderosis; C- Cholestasis (black arrows); D- Focal necrosis (blue arrow) and 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.

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 response variable 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-subject 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 will be used considering the model with the main effect given by

$$\text{logit}\left[\gamma_{ij}(x_i)\right] = \log\left[\frac{\gamma_{ij}(x_i)}{1 - \gamma_{ij}(x_i)}\right] = \alpha_j + \beta_j x_i, \quad j = 1, 2 \tag{6}$$

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 (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. 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 & Johnson, 1997)

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

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 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

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

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 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 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 & 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).

6. Results and Discussion

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 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

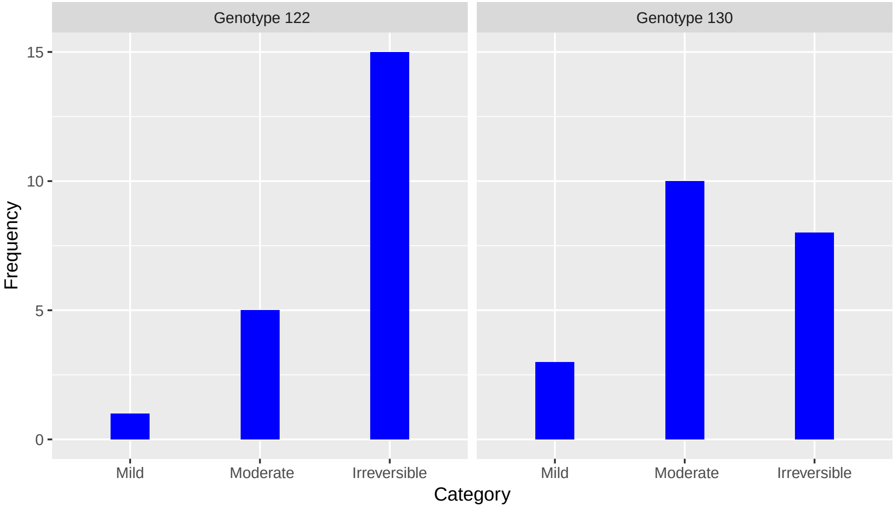

Figure 3. Frequencies of mild, moderate, and irreversible lesions in the liver of Tambaquis by type of genotype (122 and 130) in the study carried out by Marques (2018) at the Biofish-Aquicultura farm.

fish with genotype 122. 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 1.

Table 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 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. Fish with genotype 122 showed irreversible 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, the confidence interval has greater amplitude due to the relatively small sample size.

Table 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 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).

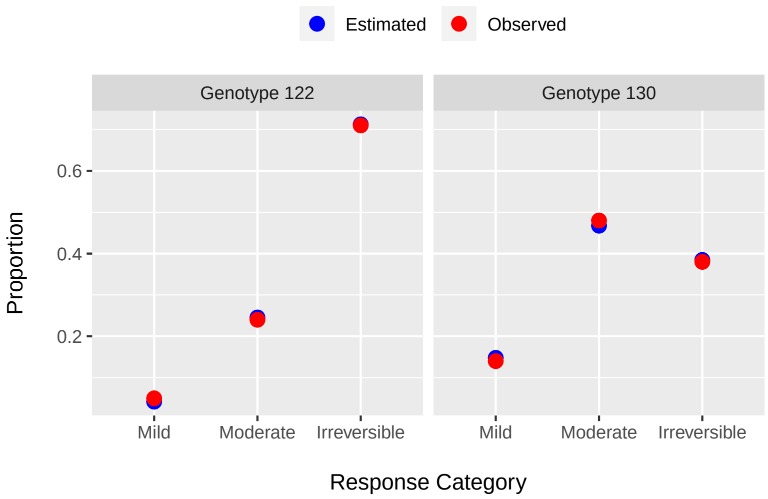

Figure 4. Observed proportions for the mild, moderate, and irreversible lesions and proportions estimated by proportional odds model with genotype effect in the study of Marques (2018).

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 5, the residual distribution presented a shape similar to the standard logistic distribution (red line), 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.

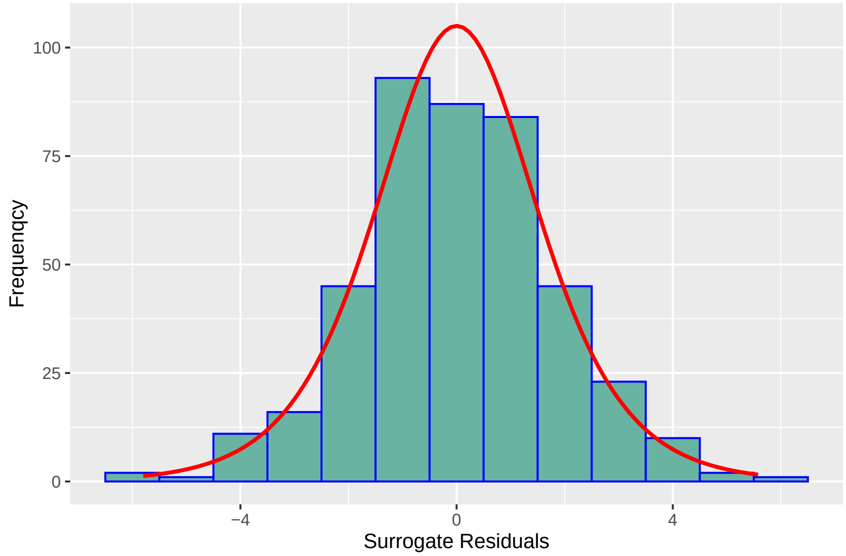

Figure 5. Histogram of surrogate residuals related to the proportional odds fitted model (genotype effect) to the fish data in the study of Marques (2018)

The half-normal plot with a simulated envelope for the surrogate residuals was presented in Figure 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.

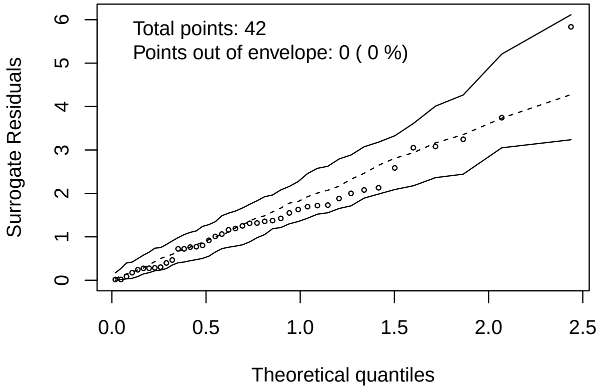

Figure 6. Half-normal plot with a simulated envelope (the default is 0.95 confidence level) for the surrogate residuals to assess the fit of the model with genotype effect in the study of Marques (2018).

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 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.

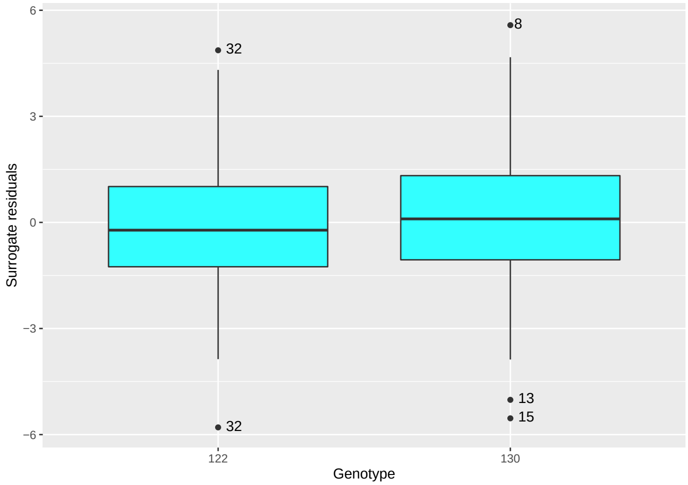

Figure 7. Boxplot of surrogate residuals per genotype to assess the proportional odds fitted model to the fish data in the study of Marques (2018).

The large residuals in the Figure 7 refer to subject #32 for genotype 122 and to subjects #8, #13, and #15 for genotype 130. The model was fitted without these subjects to assess the impact on the estimates of model parameters. The parameter estimates and the related standard errors, in parentheses, are shown in Table 3.

Table 3. Estimated Parameters of proportional odds model with genotype effect by excluding the subject #32 for genotype 122 and the subjects #8, #13 and #15 for genotype 130 the fish data

subject	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 subjects by genotype from the sample (Table 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.

7. Conclusions

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 subjects. 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 subject 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.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Agresti, A. *An introduction to categorical data analysis* 2nd ed., 394 (John Wiley & Sons, Hoboken, New Jersey, 2007).
2. Agresti, A. *Analysis of ordinal categorical data* 2nd ed., 405 (John Wiley & Sons, Nova Jersey, 2010).
3. Agresti, A. *An introduction to categorical data analysis* 3rd ed., 375 (John Wiley & Sons, Nova Jersey, 2002).
4. Ananth, C. V. & Kleinbaum, D. G. Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology* **26,** 1323–1333 (1997).
5. Arbogast, P. G. & Lin, D. Model-checking techniques for stratified case-control studies. *Statistics in medicine* **24,** 229–247 (2005).
6. Arnold, T. B. & Emerson, J. W. Nonparametric goodness-of-fit tests for discrete null distributions. *R Journal* **3** (2011).
7. Atkinson, A. C. *Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis* tech. rep. (1985).
8. Bilder, C. R. & Loughin, T. M. *Analysis of categorical data with R* 1st ed., 547 (Chapman and Hall/CRC Press, Boca Raton, 2014).
9. Christensen, R. H. B. ordinal: Regression models for ordinal data. *R package version* **28,** 56 (2013).
10. Cook, R. D. Detection of influential observation in linear regression. *Technometrics* **19,** 15–18 (1977).
11. Cook, R. D. & Weisberg, S. *Residuals and influence in regression* 248 (New York: Chapman and Hall, 1982).
12. Correa, R. O., Souza, A. R. B. & Martins Junior, H. Criação de tambaquis. *Embrapa Amazônia Oriental-Fôlder/Folheto/Cartilha (INFOTECA-E)* (2018).

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

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

15. Faraway, J. J. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models* 399 (CRC press, 2016).

16. Giolo, S. R. *Introdução à análise de dados categóricos com aplicações* 1st ed., 256 (Editora Blucher, São Paulo, 2017).

17. Greenwell, B. M., McCarthy, A., Boehmke, B. C. & Liu, D. Residuals and Diagnostics for Binary and Ordinal Regression Models: An Introduction to the sure Package. *The R Journal* **10,** 381–394 (2018).

18. Junior, D. L. & Veiga, R. D. ANÁLISE DE DIAGNÓSTICO EM MODELOS DE REGRESSÃO NORMAL E LOGÍSTICA. *Brazilian Journal of Biometrics* **38,** 449–482 (2020).

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

20. 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. & De Paula, L. V. T. Modeling the acceptability of cashew apple nectar brands using the proportional odds model.*Journal of Sensory Studies* **30,** 136–144 (2015).

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

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

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

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

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

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

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

28. McCullagh, P. & Nelder, J. *Generalized Linear Models* 2nd ed., 375 (Chapman and Hall, London, 1989).

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

30. Ng, K. W., Tian, G. L. & Tang, M. L. *Dirichlet and related distributions: Theory, methods and applications* 1st ed., 336 (John Wiley & Sons, 2011).

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

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

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

34. Reiter, J. P. & Kohnen, C. N. Categorical data regression diagnostics for remote access servers. *Journal of Statistical Computation and Simulation* **75,** 889–903 (2005).
35. Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **52,** 591–611 (1965).
36. Silva, J. A. P. *Métodos de diagnóstico em modelos logísticos trinomiais* Dissertação (Mestrado em Estatística) (Universidade de São Paulo, 2003), 100.
37. Simonoff, J. S. *Analyzing categorical data* 1st ed., 496 (Springer, 2003).
38. Singer, J. M., Rocha, F. M. M & Nobre, J. S. Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review* **85,** 290–324 (2017).
39. Souza, E. C. *Análise de influência local no modelo de regressão logística* MA thesis (ESALQ/Universidade de São Paulo, 2006), 102.
40. Turkman, M. A. A. & Silva, G. L. Modelos Lineares Generalizados–da teoria à prática. *Sociedade Portuguesa de Estatística, Lisboa,* 153 (2000).
41. Tutz, G. *Regression for categorical data* (Cambridge University Press, Cambridge, 2011).
42. Williams, O. D. & Grizzle, J. E. Analysis of contingency tables having ordered response categories. *Journal of the American Statistical Association* **67,** 55–63 (1972).

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 subjects 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)
```