DOI: 10.5433/2237-4876.2023v26n3p115-128

Introduction to the R package Ordinal for sociolinguistic evaluation analyzes

Silvia Carolina Gomes de **SOUZA GUERREIRO***
Gabriel **SALES****
Eliete Figueira **BATISTA DA SILVEIRA*****

* Doctoral degree in Portuguese Language from the Universidade Federal do Rio de Janeiro. Email address: silviacarolinasouza@gmail.com.
** Master's degree in Portuguese Language from the Universidade Federal do Rio de Janeiro. Email address: gabriel-sales@outlook.com.
*** Doctoral degree in Portuguese Language from the Universidade Federal do Rio de Janeiro. Professor at the UFRJ. Email address: elietesilveira@letras.ufrj.br.

Abstract:
This article introduces the Ordinal package (Christensen, 2022) for the analysis of social evaluation tests, highlighting its contributions to sociolinguistic research. Evaluation studies have the potential to generate data of very different natures since there is considerable variability in methodological procedures. This work focuses on ordinal data, whose levels establish a hierarchical relationship (for example, A>B>C). To illustrate the analysis, we used the corpus of Souza Guerreiro (2023), which investigates the evaluation of speakers from Rio de Janeiro about the pretonic vowel-raising phenomenon. The results indicate that the participants tend to elect elementary school to the speaker who produces a raised vowel in syllable ending with a rhotic. On the other hand, when syllables end with a sibilant, the tendency is to attribute secondary education to the speaker. The main advantages of the Ordinal package identified in the analysis were: there is no limitation of levels of the dependent

variable; the model predictions are produced in a mathematically adequate way to the ordinal nature of the data, and the model output allows to visualize, in a scalar way, the effect of independent variables on the levels of the dependent variable.

Keywords:
Ordinal; Social evaluation; Sociolinguistics;

Signum: Estudos da Linguagem, Londrina, v.26, i.3, p.115-128, december. 2023
Received on: 30/05/23
Accepted on: 04/12/23

Introduction to the R package Ordinal for sociolinguistic evaluation analyses

Silvia Carolina Gomes de Souza Guerreiro
Gabriel Sales
Eliete Figueira Batista da Silveira

INTRODUCTION

Traditionally, statistical tools used in variationist research are Binary Logistic Regression methods, whose optimal model, also called *step* in this *framework*, has its variables ordered by relevance/contribution to the explanation of the *dependent variable* (VD) through the *relative weights* generated.

Since the advent of Sociolinguistics, several efforts have been made to automate this analytical process, resulting in the distribution of widely used *software* such as *Varbrul* (Rousseau; Sankoff, 1978), *GoldVarb X* (Lawrence; Tagliamonte; Smith, 2005) and *Rbrul* (Johnson, 2009). Although it has been quite fruitful for the area, the development of such tools reflects the prioritization of linguistic production research marked in the 20th century, although the role of *social evaluation* had already been duly recognized since Weinreich, Labov, and Herzog (1968).

Research focused on exploring the subjective aspects of the linguistic system and aiming to facilitate statistical treatment of its data using specific *software* programs must adapt its data collection procedures based on the inherent *input* limitations of each tool. Typically, the dependent variable (DV) should be nominal and possess only two levels, such as "*yes*" and "*no*." Notably, *Rbrul* enables the control of random variables[1]. Despite advancements over its predecessors, it's worth noting that *Rbrul* is confined to the Binary Logistic Regression method.

Given this, it is evident that the restricted use of traditional *software*[2] can be a limiting factor for the researcher's creativity regarding the collection method and the type of data generated, thus compromising the variability of research techniques used. That demonstrates the need to expand the repertoire of statistical treatment methods within the scope of Sociolinguistics, especially in analyzing data originating from subjective reaction tests resulting from experimental methods capable of generating data of very diverse natures.

In this study, our emphasis is on ordinal data — variables with hierarchical relationships among their levels, such as A > B > C > […]. Such data typically arises from sources like Likert

[1] According to Baayen (2008), variables that do not have the property of limiting observable levels are labeled in this way. For example, in a subjective reaction test in which the *informant* variable is controlled, each individual who responds to the test will result in the addition of a new level. Furthermore, the levels constituted in this type of analysis have a low chance of replication: a new collection, with new informants, will possibly capture a different sample of the studied population, consequently recording new levels. Random variables are opposed to fixed variables, whose levels are limited and whose reproduction can be controlled (for example, structural factors concerning stimuli, such as the realization of a segment and the phonetic context). The incorporation of both types of variables is done in mixed-effects model analyses.

[2] The limitation also extends to the linguistic production research itself since the tools mentioned require the limitation of DV levels. Thus, it is not possible, for example, to analyze, in a single model, quaternary variables, such as the realizations of /S/, whose variants can be [s] ~ [ʃ] ~ [h] ~ [Ø]. The methodological process, in this case, ends up being a little more arduous, as there is a need to choose a reference to be compared to each of the other variants in 3 different analyses.

scales[3]. The inferential treatment of ordinal data necessitates the modeling of cumulative probabilities, aiming to discern the accumulated probability associated with each scale level (Garcia, 2021) and assess the probabilistic distinctions between them (Baayen, 2008). Within the R environment, numerous packages enable analyses of this nature. Nevertheless, our focus in this work is specifically on the *Ordinal* package (Christensen, 2022), which facilitates the incorporation of random variables in the analysis through the *clmm*() function.

Linguistic studies, encompassing subjective evaluations, frequently involve dealing with random variables, be it the lexical item itself or the informant. Consequently, researchers are better served by opting for tools that accommodate such coding, thereby enhancing the alignment of statistical model predictions. In this regard, using the discussed statistical package proves highly productive, particularly in contexts demanding the analysis of ordinal variables.

In this text, we intend to present the tool, its method of use, and general information for interpreting its *output*. Prior knowledge of R, although desirable, is not required to read the work since we cover inferential analysis from the initial steps in this environment. Next, we list the advantages of the statistical package discussed and its potential. Finally, we summarize the tool's contributions to the area and recommend readings for a deeper understanding of its operation and the R language.

DEFINITION OF VARIABLES

To illustrate the use of the *Ordinal* package, we modeled analyses of responses from speakers in Rio de Janeiro state to a subjective reaction test to the raised realization of pretonic front vowels. The data we use is part of the study by Souza Guerreiro (2023), in which broader and more in-depth analyses can be consulted. In this article, we start from a didactic outline of variables and levels controlled by the author since our objective is effectively to present and evaluate the contributions of the analytical tool under discussion for sociolinguistic studies.

The data were collected by Souza Guerreiro (2023) through an experiment designed according to the matched-guise technique (Lambert; Lambert, 1975). In this work, 200 informants evaluated the probable level of education (elementary, secondary, or higher) of readers of sound stimuli with the production of sentences in which the pretonic vowels of some words were produced sometimes with a mid-front vowel, sometimes raised.

The analysis of the results of this experiment developed in this work intends to investigate the influence, on the participants' responses, of the *structure of the syllable nucleated by the raised vowel* and *the frequency of application of the process in the lexical item.* Therefore, only the set of responses to stimuli with [i] production of the pretonic vowel is assumed as VD. The independent variables (IV) controlled are the *syllabic structures* 1. closed by sibilant ([i]studo, [i]scola), 2. by rhotic (s[i]rviço, t[i]rceira), and 3. by nasal ([i]nsino, [i]mpresa). The *frequency of occurrence of the vowel* [i] in the lexical item in the Nurc-RJ and Concordância *corpora* is also controlled, involving items with *high frequency* of raising (m[i]nino, p[i]rigo) and with *low frequency* (acad[i]mia, v[i]stibular). The relevance of the structural factors mentioned is hypothesized by Souza Guerreiro (2023) based on previous literature on pretonic raising in the speech of Rio de Janeiro (Avelheda, 2013; Souza, 2017). In section 6, where we discuss the interpretation of the results we arrived at, the work of Souza Guerreiro (2023) is discussed in greater detail.

PREPARING THE FILE FOR IMPORT

[3] Data collection method proposed by Likert (1932) for analyzing attitudes and opinions.

When performing data manipulation in R, the RStudio[4] (2022) interface is employed. Before importing and reading data in this environment, it is crucial to organize it into a specific structure, aligning with the input format used by packages like *Rbrul*. This structure comprises columns representing individual variables, with rows corresponding to their respective observations or occurrences. The first row of each column typically serves as the variable's name. Subsequent rows record the occurrences. Chart 1 below provides a visual representation of the described structure.

Chart 1 – data file structure

DV	IV 1	IV 2	IV 3	IV *n*
Observation 1	Observation 1	Observation 1	Observation 1	Observation 1
Observation 2	Observation 2	Observation 2	Observation 2	Observation 2
Observation *n*	Observation *n*	Observation *n*	Observation *n*	Observation *n*

Source: prepared by the authors

Compilation of the data file can be done directly in spreadsheet editors such as *Calc (LibreOffice)* and *Excel (Microsoft Office)* or exported in the desired format from other *software*, such as *Elan*. To read the spreadsheet in R, we use the .CSV UTF -8 format, although the *software* allows, with different functions, the import of data in other formats.

IMPORTING AND PREPARING DATA IN RSTUDIO

The first step in any data manipulation in RStudio is defining the working directory, i.e., the folder on the computer where the files we intend to work with are stored. This process can be done manually, by clicking on *Go to directory*, in the lower right *Files* tab, and navigating to the desired folder. After locating and selecting the folder, the files in it should be available to view in *Files*. So, we click on the gear icon ⚙ in that same tab and select *Set as working directory* [5].

Following this procedure, it is essential to install the packages that will be utilized. This step is a one-time requirement and can be accomplished using the *install.packages()* function. In the R language, functions are coded sequences that execute specific commands. Users typically input the parameters for these commands within parentheses (). For the installation of specific packages, such as *Tidyverse* and *Ordinal* in this instance, their names must be enclosed in double quotes within parentheses, as illustrated in (1). Once installed, these packages need to be loaded using the *library()* function, as shown in (2), ensuring accessibility to R throughout the current session. Unlike the installation, this step must be reiterated with each new session of the tool.

Chart 2 – import and preparation *script*[6]

[4] Download links for R and RStudio are provided in the references section.

[5] Once the working directory is defined by this method, the command line corresponding to this action should appear in the lower left *Console* screen. It is recommended that this same code be added to the *script* to ensure faster resumptions in any new sessions in RStudio.

[6] In the code, “fundamental”, “medio” and “superior” correspond, in Portuguese, to “ensino fundamental”, “ensino médio” e “ensino superior”, which are the levels of education in Brazil, respectively *elementary school*, *high school* and *higher education*.

```
# (1) Install packages
install.packages("tidyverse")
install.packages("ordinal")

# (2) Load packages
library(tidyverse)
library(ordinal)

# (3) Import data:
esc.pretonicas = read_csv2("dados.csv") %>%
        mutate_if(is.character, as.factor)

# (4) Inform the ordinal nature of the VD and adjust the order of the levels
esc.pretonicas.ord = esc.pretonicas %>%
        mutate(VD = factor(VD,
                        levels = c("fundamental", "medio", "superior"),
                        ordered = TRUE))
```

Source: prepared by the authors based on Garcia (2021)

Once the packages have been loaded, we are finally able to import the data file with the *read_csv2()* function, part of the *Tidyverse* package. Using the "=" operator, we assign a name to our data file, with which we will identify it as an object in the RStudio interface. In the example in (3), we chose to call our object **esc.pretonicas**. The "=" operator indicates that **esc.pretonicas** must correspond to the result of the operation applied by the *read_csv2()* function, which takes as an argument the name of the **dados.csv** file stored in our working directory.

The pipe operator "%>%" is applied to the result of *read_csv2()*, which, fundamentally, enables a chain of actions. In this case, its practical effect is to take the object loaded by *read_csv2()* and apply the *mutate_if()* function, which, in turn, promotes transformations in the data file variables according to the specified arguments: if, in the file, there is some variable of type *character*, this must be converted into a variable of type *factor* [7].

When the line of code in (3) is executed, the **esc.pretonicas** object is made available in the upper right *Environment* tab. Now, as a preparatory step, all that remains is to inform R that our DV presents an ordinal relationship between its levels. To do this, as exemplified by (4), we create a new object, which we call **esc.pretonicas.ord**, corresponding to the result of applying the *mutate()* function to **esc.pretonicas** (remember the chaining relationship promoted by the pipe). Unlike *mutate_if(), mutate()* has no conditional application: this function is invariably applied to a specific column of **esc.pretonicas**: DV. With the syntax of the *mutate()* function presented in (4), we establish that, in the new **esc.pretonicas.ord** object, the DV column must be treated as a variable of type *factor*, whose levels are, from smallest to largest, *" fundamental*", *"medium"* and *"superior*" and that this ordering is part of the nature of the variable (which is why we mark *ordered* = TRUE).

MODEL BUILDING

With the file loaded and adjusted, we move on to the analysis stage. Naturally, every inferential analysis must be preceded by an exhaustive exploratory analysis by the researcher, in

[7] *Character* and *factor* are some classes of variables in R language. The first treats observations as textual occurrences. The consequence of such treatment is that identical observations are recognized as distinct, that is, there is no establishment of "levels" of variables by grouping. The second class allows identical observations to be counted as different occurrences of the same level of a categorical variable.

order to 1. check the distribution of data according to the controlled IV and 2. derive, by comparing the distributional patterns with the available literature, hypotheses that support the inferential models to be tested. Methods for promoting exploratory analysis in R can be consulted in the specialized manuals recommended at the end of this article. The descriptions presented here cover only inferential analysis procedures and presuppose prior investigation of data distribution.

Chart 3 – Basic syntax of ordinal mixed-effects models[8]

```
# (5) Generate model 1
mod1 = clmm(VD ~ FREQUENCIA + ESTR.SILABICA +
                (1|INFORMANTE), data = esc.pretonicas.ord)
```

Source: prepared by the authors

Chart 3 presents the basic structure of an ordinal model built with the *clmm()* function. The structure does not differ much from that used in more popular functions, such as *glm():* the first argument is the DV – the variable whose behavior we intend to explain. This property is encoded by "~", which represents the limit between the DV and the following argument(s): the IV, possible factors conditioning the variability of the DV. There is no limitation on the number of independent variables coded, nor on their nature (quantitative or qualitative)[9]. In Chart 3, as we deal with more than one IV, we establish the interleaving between them using "+". This symbol, however, considers the isolated action of each IV on the DV. In addition to this scenario, it is possible for two IV to act together, configuring an interaction. To encode this analytical possibility in R, we replace "+" with "*****".

If the research includes any random variable – in our case, the informant –, this must be introduced by "+", following the notation (1|**VAR**) assumed by the package, in which, to the right of the vertical bar, it is specified the randomness factor. Finally, with the last argument, *data* = **esc.pretonicas.ord**, we specify, to the right of "=", the R object in which the data to be modeled is contained.

COMPARISON BETWEEN MODELS

The outcome of the syntax provided in the preceding section is a model where evaluators' responses to the fundamental question in the research instrument (*what is the likely educational level of this individual?*) are elucidated by the isolated action of the variable's *frequency* of raising the pretonic vowel in the lexical item and *syllabic structure*. Following the execution of code (5), the obtained results and corresponding statistics can be reviewed and interpreted in the *console* using the *summary()* function. However, it's crucial to note that the process of model generation doesn't conclude at this stage. The researcher must assess the significance of incorporating one variable over another and evaluate the impact of potential interactions hypothesized based on the cross-distribution of data, contributing to the model's explanatory power.

This type of analysis is conducted, in the case of ordinal models, by *likelihood ratio tests*, which can be performed with the *anova()* function. To demonstrate this process, first, we built a second model, with code (6) from Chart 4, in which we started to consider a possible interaction

[8] In the code, "frequencia", "estr.silabica" and "informante" refer to the Portuguese words "frequência" "estrutura silábica" and "informante", respectively meaning, "frequency", "syllabic structure" and "informant".

[9] In any probabilistic program, the amount of data can be a limitation, considering that a very small sample has the potential to generate less reliable results, especially when associated with a large number of VI.

between our IV. Subsequently, with code (7), we establish an effective comparison between the two models.

Chart 4 – comparison between fixed and mixed-effects models

```
# (6) Generate model 2 (mixed effects)
mod2 = clmm(VD ~ FREQUENCIA * ESTR.SILABICA +
                (1|INFORMANTE), data = esc.pretonicas.ord)

# (7) Compare models
anova(mod1, mod2)
```

Source: prepared by the authors based on Christensen (2022)

The *anova* function in (7) takes the two constructed models as arguments. Its result, in all applications, should be similar to that presented in Chart 5, which reproduces the *console output*.

Chart 5 – *output* of *anova function*

```
Likelihood ratio tests of cumulative link models:

      formula:                                          link: threshold:
mod   VD ~ FREQUENCIA * ESTR.SILABICA + (1 | INFORMANTE) logit flexible
2
mod   VD ~ FREQUENCIA+ ESTR.SILABICA + (1 | INFORMANTE) logit flexible
1

      no.par  AIC     logLik    LR.sta   df  Pr(>Chisq)
mod   6       515.46  -251.73
2
mod   7       515.28  -250.64   2.179    1   0.1399
1
```

Source: prepared by the authors

The first line of the *output* informs the type of test performed. Below this information, the formulas for each of the compared models are summarized. Finally, we have access to the quality statistics of each model. In short, the *anova()* function evaluates the differences between the compared elements. If there is no significant distinction between the two, the test will *output* a *p-value* higher than the significance level ($\alpha = 0.05$). However, if there is a significant difference, the test will return a *p-value* lower than the α level.

The goal of inferential analysis is to achieve the greatest explanatory power through the simplest modeling. As a consequence of this, when comparing models, when $p>0.05$, we opt for the simplest among those compared: the one that contains fewer variables or, in our case, the one that does not encode interaction between predictors. However, when the *output* of the *anova* function displays $p<0.05$, the significance of the difference between the models prevents us from choosing the simplest one. In this situation, the most complex option must be assumed to be optimal modeling.

INTERPRETATION OF RESULTS

Before discussing the interpretative details of the model selected in the previous section, we consider it relevant to revisit some details and hypotheses from the research by Souza Guerreiro (2023), in order to contextualize the data we deal with and the objectives we intend to achieve with ordinal logistic regression analysis.

In Souza Guerreiro's research (2023), participants were exposed to nine audio stimuli featuring a raised pretonic vowel. The primary objective of the study was to observe the informants' tendencies in attributing a *level of education* to the speaker exhibiting the raised pretonic vowel. After the audio exposure, participants responded to the query: "*If you had to assign a level of education to this speaker, you would say that he/she attended: (i) elementary school I, former primary school; (ii) elementary school II, former gymnasium; (iii) high school; (iv) higher education.*" To analyze the outcomes, the variables "*elementary education I, former primary*" and "*elementary education II, former gymnasium*" were merged, resulting in a dependent variable with three levels. This analytical approach aimed to discern whether participants tended to attribute lower education levels to speakers who raised the pretonic vowel.

In this work, with the didactic objective of presenting the analytical tool, we observe the importance of only the *syllable structure* and *frequency* for evaluating pretonic raising. Specifically, speakers' assessments are analyzed about the raising of the front vowel in syllables closed by a sibilant – e.g., [i]scola (*escola,* meaning school), [i]squina (*esquina*, meaning street corner) – and by a rhotic – e.g., s[i]rviço (*serviço*, meaning service), s[i]rvidos (*servidos*, meaning served).

Previous production studies (Avelheda, 2013; Souza, 2017), when analyzing the syllabic context, observed, from a *continuum,* that the raising of the front vowel tends to occur in syllabic structures closed by a sibilant – e.g., [i]scola (*escola,* meaning school), [i]studo (*estudo*, meaning study) – and a nasal sound – e.g., [i]mpregada (*empregada*, meaning house maid), [i]nsino (*ensino*, meaning teaching). However, the researchers found that vowel raising is rarely recurrent in the structure of free syllables – e.g., s[e]nhor (*senhor*, meaning sir), p[e]queno (*pequeno*, meaning small) – and syllables closed by rhotic – e.g., s[e]rviço (*serviço*, meaning service), t[e]rceira (*terceira*, meaning third).

Therefore, considering the results of the production research, the hypothesis was created that the informants would evaluate: (i) positively the speaker who produces the raising in a syllable closed by a sibilant, as the phenomenon is recurrent in this syllabic structure and (ii) negatively the speaker who produces the raising vowel in a rhotic-ending syllable, as the phenomenon tends not to be produced in this syllabic context. Such hypotheses exactly include aspects concerning the *structure of the syllable* and the *frequency of raising*, justifying their inclusion in the previously exemplified regression models.

Armed with this information, code (6) in Chart 4, corresponding to our optimal model, can now be resumed. To visualize the results, we use the *summary()* function, taking our model as an argument, as exemplified by code (8) in Chart 6, whose *output*, reproduced in Chart 7, are the results of the regression analysis.

```
# (8) Request results to be displayed in the console
summary(mod2)
```

Chart 6 – model results visualization
Source: prepared by the authors

Chart 7 – partial *output* of the *summary()* function in the console

Coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
FREQUENCYlow	0.3610	0.2662	1.356	0.1751
ESTR.SILABICAt.r	-1.9385	0.4133	-4.690	2.73e-06 ***
ESTR.SILABICAt.s	0.5659	0.2681	2.111	0.0348 *

Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			
Threshold coefficients:				
	Estimate	Std. Error	z value	
elementary | high school	-0.9449	0.2546	-3.711	
High school | higher education	0.9722	0.2557	3.803	

Source: prepared by the authors

The *output* in Chart 7 is originally preceded by some information: the identification of the type of regression, the model formula, the data set, and the specification of the random variables. To preserve space, however, we chose not to reproduce them. Observing Chart 7 highlights two sets of results. The first, *coefficients*, presents statistics referring to each level of the model's independent variables, except reference levels. As the *frequency* variable is binary, just one comparison is enough to account for the opposition between its levels. The estimate displayed, therefore, corresponds to the opposition between *low and high frequency* when the other variables in the model are at their respective reference levels. On the other hand, the *syllabic structure* variable is ternary. Therefore, two comparisons are necessary to cover all possible oppositions of syllable closing: one that contrasts nasal (i.e., the reference level) with rhotic and another that contrasts the reference with sibilant.

The estimates generated for each model's oppositions are measured in *logodds*, a unit centered on 0. Generally, in conventional logistic regression, values in *logodds* above 0 indicate favorability, and below 0, disfavor. In the case of ordinal logistic regression, however, the estimates must be read in terms of the second group of results, the *Threshold coefficients*. *Thresholds* indicate numerically, in *logodds*, the limits between the levels of the dependent variable analyzed. By comparing the coefficients of the independent variables and those of limits, it is, therefore, possible to identify which of the DV levels tends to be most likely attributed according to the variation of the different predictors.

Let us see, in practice, based on the data analyzed, what this relationship between IV coefficients and limit coefficients means. Once the *syllable structure* and *frequency* variables were submitted to the model, the independent variables *syllable closed by rhotic* and *syllable closed by sibilant* were selected as significant.

To understand the analysis, it is presented how the results are read. After selecting the most significant round, the following threshold scale (*Threshold coefficients*) measured in *logodds* for the levels of the dependent variable *education* is extracted from Chart 7:

Table 1 – threshold coefficients

Upgrade assessment – Education – Threshold coefficient[10]
Elementary School : < -0.95 logodds
High School: -0.95 logodds to 0.97 logodds
Higher Education: > 0.97 logodds

Source: prepared by the authors

[10] When reproducing the results in logodds, only two decimal places were considered.

According to the results above, it is understood that: (i) if the estimate of a IV is <-0.95 *logodds*, there is a greater probability that the informants will choose, at least, elementary school education for the speaker of the raised variant ; (ii) if the estimate is between -0.95 and 0.97 *logodds*, informants tend to select secondary education and (iii), if the estimate is > 0.97 *logodds*, there is a greater probability of informants selecting higher education. Based on this guidance, we proceed to read the results of estimates of the independent variables, extracted from Chart 7.

Table 2 – model report

Upgrading assessment - Education			
Elementary School: < -0.95 *logodds*			
High school: -0.95 a 0.97 *logodds*			
Higher Education: > 0.97 *logodds*			
Selected variables	**Estimates**	**P. value**	**Minimum education assigned**
Syllable closed by /R/	-1.94 *logodds*	$2.73e^{-06}$	Elementary School
Syllable closed by /S/	0.57 *logodds*	0.03	High school

Source: prepared by the authors

When analyzing the results of the rhotic-closed syllable variable, it is observed that the estimate -1.94 *logodds* is included in the results relating to elementary school (< -0.95 *logodds*). Therefore, according to the result, if the raised front vowel is in a rhotic-closed syllable structure, such as s[i]rviço (*serviço*, meaning service), informants tend to choose *elementary school* education for the speaker.

Concerning the variable *syllable closed by sibilant*, the estimate of 0.57 *logodds* is found among the results referring to *secondary education* (between -0.95 and 0.97 *logodds*). Thus, the result indicates that informants tend to attribute *high school education* to the speaker who raises the previous vowel in a syllable closed by a sibilant, such as [i]scola (*escola*, meaning school).

Therefore, the results of the perception study confirm the hypothesis raised by Souza Guerreiro (2023) that informants tend to attribute *low education* to the speaker who raises the front vowel followed by rhotic (e.g., s[i]rviço, s [i]rvidos, t[i]rceira), possibly because the phenomenon does not tend to be produced in this syllabic context. However, informants tend to attribute high school education to the speaker who raises the previous vowel in a syllable closed by a sibilant (e.g., [i]studo, [i]scola, [i]scritório). That indicates that, in the context in which the heightened realization is often produced, its realization does not seem socially salient (Avelheda, 2013; Souza, 2017)[11].

ASSESSMENT OF THE TOOL'S POTENTIAL

The inferential analysis of ordinal data through the *clmm()* function of the *Ordinal* package (Christensen, 2022) has the potential to contribute to the methodological refinement of subjective evaluation research within the scope of Sociolinguistics, as it enables the adequate treatment of variables whose levels establish some hierarchical relationship. Among the advantages of its use, we list the following:

1. There is no limitation on the levels of the dependent variable, which represents an important advantage in relation to the strict use of binary logistic regression. Ordinal regression, however, is not necessarily a substitute for binary since both methods serve

[11] Considering the categories of Labov (2001 [1966]; 2008 [1972]), *indicator*, *marker* and *stereotype*, Souza Guerreiro (2023) interprets, based on these results, that the vowel raising in syllables ended in /S/ is a *marker*. In a rhotic-closed syllable structure, the raising is a *stereotype*. About this, read Souza Guerreiro (2023).

data of different natures. The expansion of the repertoire of analytical methods, however, makes it possible to develop projects with different methodological designs that generate different types of data and answer different research questions;

2. There is no limitation on the nature of the independent variables, as there is, for example, when using *Varbrul* or *GoldVarb X*. As it is based on R software, *Rbrul* also provides this benefit. Therefore, with *Ordinal* and *Rbrul*, it is possible to incorporate continuous variables, such as the age of individuals, without the mandatory need to discretize them, segmenting them into age groups. In this way, the analyst has greater flexibility regarding how their data is processed, which must be chosen depending on the research objectives;
3. It is possible to construct mixed effects models. That is an advantage shared among many tools built on R, including *Rbrul* and *Ordinal*. Ideally, random variables should be included in the model whenever they make up the research "variation package", in order to confirm whether the effects of the fixed variables are maintained (Oushiro, 2022);
4. By considering cumulative probabilities, the model's predictions are produced in a manner that is mathematically appropriate to the nature of the data, respecting the ordering between the levels of the dependent variable, and
5. the results *output* reflects in a very intuitive way the trends of change in forecasts according to the independent variable analyzed. In this way, it is possible to visualize its probabilistic distribution on the hierarchical scale established between the levels of the dependent variable.

CONCLUSION

In Sociolinguistics, subjective assessment tests have different objectives and use different methodological designs. As a result, data of different types are produced that require analysis based on their specific characteristics. This article investigates an ordinal data analysis tool, witch can be used for analyzing data resulting from some sociolinguistic perception tests. The *ordinal* package provides appropriate statistical resources for studying ordinal variables in this context. It provides probabilistic results that are calculated using the hierarchy between the individual levels, as seen in the example of the *education* variable. The levels of this variable are systematically arranged as follows: *Primary education* < *Secondary education* < *Higher education*.

The use of this method made it possible to visualize, based on data from Souza Guerreiro (2023), that informants tend to evaluate the speaker of the raised variant [i] followed by rhotic – s[i]rviço (*serviço*, meaning service) – as a person with a low level of education. There is, therefore, a specific social labeling of this speaker. About the syllable closed by sibilant – [i]scola (*escola*, meaning school), participants tend to attribute high school education. This result is adequate because the model considers the hierarchy between the DV levels when calculating its probabilistic predictions.

If hierarchy is a determining factor in the tool adopted for the data analyzed, in other cases, it may not be. For this reason, the researcher must be aware of the nature of the data generated since its characteristics determine the most appropriate analysis method. Because of this, different analytical tools should be combined into the analyst's repertoire so that he or she can answer his or her research questions without this being a limiting factor in the methodological designs used.

We hope this work contributes to disseminating an inferential method that applies to the analysis of perception studies. The topic, however, does not end in this article. For this reason, other readings are recommended. Christensen (2022), author of the *Ordinal* Statistical Package, exhaustively documents the functions that make up the tool. The *clm()* function, from the same

package, is explored by Garcia (2021), who describes how it is used. This function is analogous to *clmm()* but is only suitable for analyzing fixed variables. As we did here, the construction of mixed models with the Ordinal package is exemplified in detail by Barlaz (2022) on her personal website. Baayen (2008), in turn, describes, in addition to ordinal regression, linear and logistic regression methods. The author also dedicates a particular chapter to the discussion of mixed models. Levshina (2015) also addresses various uni and multivariate analysis methods. In Portuguese, there are works by Oushiro (2015, 2022) and Gries (2019), in addition to the courses *Quantitative Data Analysis in Linguistics* (Lima Jr., 2021) and *Statistics for Language Sciences* (Godoy, 2021), available on public *YouTube* playlists. There are also, on the EAD platform of the Brazilian Linguistics Association (Abralin), the courses *Regression Models for Linguists* (Lima Jr.; Garcia; Angele, 2020) and *Introduction to Statistics with R* (Oushiro, 2021).

Therefore, the interface between Linguistics and Statistics tends to collaborate with improving and adapting analytical processes. New analysis methods and updates to existing ones always appear in the statistical field. Therefore, interaction between the two areas must be constantly sought to evaluate the relevance of incorporating new methodologies in our analyses. That does not mean, however, that Statistics is the end of the work, but a *means*: a method to contribute to the processing and interpretation of data, from which reflections on linguistic knowledge are derived.

REFERENCES

AVELHEDA, A. C. C. **O alteamento das vogais médias pretônicas no município de Nova Iguaçu**: análises sociolinguística e acústica. Dissertação (Mestrado em Letras Vernáculas) – Faculdade de Letras, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013.

BAAYEN, R. H. **Analyzing linguistic data**: a pratical introduction to Statistics using R. New York: Cambridge University Press, 2008.

BARLAZ, M. **Ordinal logistic regression in R**. 2022. Disponível em: https://marissabarlaz.github.io/portfolio/ols/. Acesso em: 27 maio 2023.

CHRISTENSEN, R. H. B. **Ordinal**: regression models for ordinal data. 2022. Pacote R versão 2019.12-10. Disponível em: https://CRAN.R-project.org/package=ordinal.

GARCIA, G. D. **Data visualization and analysis in second language research**. New York: Routledge, 2021.

GODOY, M. **Estatística para as ciências da linguagem**. 2021 Disponível em: https://youtube.com/playlist?list=PLE4HwfVNrSWQwm_62G49CZTXi7dqMzsuC. Acesso em: 27 maio 2023.

GRIES, S. Th. **Estatística com R para a Linguística**: uma introdução prática. Belo Horizonte: FALE/UFMG, 2019.

JOHNSON, D. E. Getting off the GoldVarb Standard: introducing Rbrul for mixed-effects variable rule analysis. **Language and Linguistics Compass**, v. 3, n. 1, 2009. p. 359–383.

LABOV, W. **Principles of linguistic change**: Social Factors. Massachusetts: Blackwell Publishers, 2001 [1966].

LABOV, W. **Padrões sociolinguísticos**. São Paulo: Parábola Editorial, 2008 [1972].

LAMBERT, W; LAMBERT, W. **Psicologia Social**. Trad. Dante Moreira Leite. Rio de Janeiro: Zahar, 1975.

SANKOFF, D.; TAGLIAMONTE, S. A.; SMITH, E. **Goldvarb X**: a variable rule application for Macintosh and Windows. Department of Linguistics, University of Toronto, 2005.

LEVSHINA, N. **How to do Linguistics with R:** data exploration and statistical analysis. Amsterdam: John Benjamins Publishing Company, 2015.

LIKERT, R. A technique for the measurement of atitudes. **Archives of Psychology**, v. 20, n. 140, 1932.

LIMA JR., R. **Análise quantitativa de dados em Linguística**. 2021. Disponível em: https://youtube.com/playlist?list=PLzkA7H-mNfYhdbUe1e0FMpJDdLj1585zq. Acesso em: 27 maio 2023.

LIMA JR., R.; GARCIA, G. D.; ANGELE, B. **Modelos de regressão para linguistas**. 2020. Disponível em: https://ead.abralin.org/course/view.php?id=10. Acesso em: 27 maio 2023.

OUSHIRO, L. **Identidade na pluralidade:** avaliação, produção e percepção linguística na cidade de São Paulo. 2015. Tese (Doutorado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2015.

OUSHIRO, L. **Introdução à estatística com R**. 2021. Disponível em: https://ead.abralin.org/course/view.php?id=26. Acesso em: 27 maio 2023.

OUSHIRO, L. **Introdução à estatística para linguistas**. Campinas: Editora da Abralin, 2022.

POSIT TEAM. **RStudio**: integrated development environment for R. Versão 2023.3.1.446. Boston, MA: Posit Software, PBC, 2023. Disponível em: http://www.posit.co/. Acesso em: 27 maio 2023.

R CORE TEAM. **R:** a language and environment for statistical computing. Versão 4.2.1. Vienna, Austria: R Foundation for Statistical Computing, 2022. Disponível em: https://www.R-project.org/. Acesso em: 27 maio 2023.

ROUSSEAU, P.; SANKOFF, D. Advances in variable rule methodology. In: ROUSSEAU, P.; SANKOFF, D. Linguistic variation: models and methods. New York: Academic Press. 1978. p. 57-69.

SOUZA GUERREIRO, S. C. **Alteamento das vogais médias pretônicas no município do Rio de Janeiro:** décadas de 70, 90 e 2010 / estudo de crenças e atitudes. 2017. Dissertação (Mestrado em Letras Vernáculas) – Faculdade de Letras, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2017.

SOUZA GUERREIRO, S. C. **Avaliação subjetiva e indexação social do alteamento pretônico**. 2023. Tese (Doutorado em Letras Vernáculas) – Faculdade de Letras, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2023.

WEINREICH, U.; LABOV, W.; HERZOG, M. **Fundamentos empíricos para uma teoria da mudança linguística**. São Paulo: Parábola Editorial, [1968] 2006.