DOI: 10.5433/2237-4876.2023v26n3p115-128

Introdução ao pacote em R Ordinal para análises de avaliação sociolinguística

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

*Doutora em Letras Vernáculas (Língua Portuguesa) pela Universidade Federal do Rio de Janeiro (2023). Contato: silviacarolinasouza@gmail.com.
**Mestre em Letras Vernáculas (Língua Portuguesa) pela Universidade Federal do Rio de Janeiro (2023). Contato: gabriel-sales@outlook.com.
***Doutora em Letras Vernáculas (Língua Portuguesa) pela Universidade Federal do Rio de Janeiro (2003). Docente da UFRJ. Contato: elietesilveira@letras.ufrj.br.

Resumo:
Este artigo objetiva introduzir o pacote *Ordinal* (Christensen, 2022) para análise de testes de avaliação subjetiva, destacando suas contribuições para pesquisas sociolinguísticas. Os estudos de avaliação têm potencial de gerar dados de naturezas muito diversas, uma vez que há grande variabilidade de procedimentos metodológicos. O foco deste trabalho recai sobre dados ordinais, cujos níveis estabelecem relação hierárquica (por exemplo, A>B>C). Para ilustrar a análise, utilizamos o corpus de Souza Guerreiro (2023), que investiga a avaliação de falantes do Rio de Janeiro em relação ao fenômeno do alteamento pretônico. Os resultados indicam que os participantes tendem a atribuir o ensino fundamental ao falante que realiza o alteamento em sílaba travada por rótico. Já em sílaba travada por sibilante, a tendência é de atribuição de ensino médio ao falante. As principais vantagens da ferramenta identificadas na análise foram: não há limitação de níveis da variável dependente; as previsões do modelo são produzidas de maneira matematicamente adequada à natureza ordinal dos dados e o *output* do modelo permite visualizar, de maneira escalar, o efeito de variáveis independentes sobre os níveis da variável dependente.

Palavras-chave:
Ordinal; Avaliação subjetiva; Sociolinguística.

Signum: Estudos da Linguagem, Londrina, v.26, n.3, p.115-128, dezembro. 2023
Recebido em: 30/05/23
Aceito em: 04/12/24

Introdução ao pacote *Ordinal* para análises de avaliação sociolinguística

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

INTRODUÇÃO

Tradicionalmente, ferramentas estatísticas empregadas na pesquisa variacionista são métodos de Regressão Logística Binária, cujo modelo ótimo, também chamado, neste *framework,* de *rodada*, tem suas variáveis ordenadas por relevância/contribuição para a explicação da variável dependente (VD), por meio dos *pesos relativos* gerados.

Desde o advento da Sociolinguística, esforços diversos têm sido feitos para automatizar esse processo analítico, resultando na distribuição de *softwares* amplamente utilizados como *Varbrul* (Rousseau; Sankoff, 1978), *GoldVarb X* (Sankoff; Tagliamonte; Smith, 2005) e *Rbrul* (Johnson, 2009). O desenvolvimento de tais ferramentas, embora tenha sido bastante frutífero para a área, reflete a priorização de pesquisas de produção linguística marcada no século XX, embora o papel da *avaliação social* já tivesse sido devidamente reconhecido desde Weinreich, Labov e Herzog ([1968] 2006).

Pesquisas que se dediquem ao estudo dos aspectos subjetivos do sistema e que pretendam promover um tratamento estatístico de seus dados por meio de um dos *softwares* citados precisam, necessariamente, ajustar seus procedimentos de coleta em função das limitações de *inputs* inerentes a cada ferramenta. No geral, a VD deve ser nominal e apresentar apenas dois níveis (por exemplo, *sim* e *não*). O *Rbrul*, especificamente, permite, ainda, o controle do efeito de variáveis aleatórias[1]. Contudo, embora apresente avanços em relação a seus antecessores, o *Rbrul* ainda é circunscrito ao método de Regressão Logística Binária.

Diante disso, é evidenciado que o uso restrito dos *softwares* tradicionais[2] pode ser um fator limitante da criatividade do pesquisador, no que diz respeito ao método de coleta e ao tipo de dado gerado, comprometendo, assim, a variabilidade de técnicas de pesquisa empregadas. É demonstrada, dessa forma, a necessidade de ampliação do repertório de métodos de tratamento estatístico no âmbito da Sociolinguística, sobretudo no que diz respeito às análises de dados originados por testes de reação subjetiva, resultantes de métodos experimentais capazes de gerar dados de naturezas muito diversas.

---

[1] Conforme Baayen (2008), são assim rotuladas variáveis que não possuem a propriedade de limitação de níveis observáveis. Por exemplo, em um teste de reação subjetiva em que seja controlada a variável *informante*, cada indivíduo que responder ao teste implicará o acréscimo de um novo nível. Além disso, os níveis constituídos nesse tipo de análise têm baixa chance de replicação: uma nova coleta, com novos informantes, possivelmente capturará uma amostra diferente da população estudada, consequentemente registrando novos níveis. Variáveis aleatórias são opostas às variáveis fixas, cujos níveis são limitados e cuja reprodução pode ser controlada (por exemplo, fatores estruturais concernentes a estímulos, como a realização de um segmento e o contexto fonético). A incorporação de ambos os tipos de variáveis é feita em análises de modelos de efeitos mistos.

[2] A limitação se estende, ainda, às próprias pesquisas de produção, uma vez que as ferramentas citadas compelem à limitação de níveis de VD. Assim, não é possível, por exemplo, analisar, em um único modelo, variáveis quaternárias, como as realizações de /S/, cujas variantes podem ser [s] ~ [ʃ] ~ [h] ~ [Ø]. O processo metodológico, nesse caso, acaba sendo um pouco mais árduo, por haver necessidade de escolher uma referência a ser comparada a cada uma das demais variantes em 3 análises diferentes.

Neste trabalho, nosso foco recai sobre dados ordinais: variáveis cujos níveis apresentam algum tipo de relação hierárquica, como A > B > C > […]. Esse tipo de dado é gerado, por exemplo, a partir de escalas tipo Likert[3]. Seu tratamento inferencial requer a modelagem de probabilidades cumulativas, uma vez que desejamos identificar a probabilidade acumulada de cada nível de uma escala (Garcia, 2021), avaliando as diferenças probabilísticas entre eles (Baayen, 2008). No ambiente R, análises desse gênero podem ser realizadas por intermédio de diversos pacotes diferentes. No entanto, neste trabalho, abordamos especificamente o pacote *Ordinal* (Christensen, 2022), por possibilitar a incorporação de variáveis aleatórias na análise, com a função *clmm()*.

Estudos linguísticos, incluindo os de avaliação subjetiva, lidam, na maior parte dos casos, com variáveis aleatórias, como o item lexical ou mesmo o próprio informante. Sendo assim, o pesquisador deve, preferencialmente, adotar ferramentas que permitam esse tipo de codificação, em favor da adequação das previsões dos modelos estatísticos gerados. Por esse motivo, o uso do pacote estatístico em discussão pode ser bastante produtivo em contextos nos quais a análise de variáveis ordinais seja requisitada.

Neste texto, pretendemos apresentar a ferramenta, o seu método de utilização e as informações gerais de interpretação de seu *output*. O domínio prévio de R, embora desejável, não é requisitado para a leitura do trabalho, uma vez que cobrimos a análise inferencial desde os passos iniciais nesse ambiente. Em seguida, elencamos as vantagens do pacote estatístico abordado e suas potencialidades. Por fim, sumarizamos as contribuições da ferramenta para a área e indicamos leituras para aprofundamento sobre seu funcionamento e sobre a linguagem R.

DEFINIÇÃO DE VARIÁVEIS

Para exemplificação do uso do pacote *Ordinal*, modelamos análises de respostas de falantes do estado do Rio de Janeiro a um teste de reação subjetiva à realização alteada de vogais pretônicas de série anterior. Os dados de que partimos integram o estudo de Souza Guerreiro (2023), no qual podem ser consultadas análises mais amplas e aprofundadas. Neste artigo, partimos de um recorte didático de variáveis e de níveis controlados pela autora, uma vez que nosso objetivo é, efetivamente, apresentar e avaliar as contribuições da ferramenta analítica em discussão para estudos sociolinguísticos.

Os dados foram coletados por Souza Guerreiro (2023) por meio de experimento desenhado de acordo com a técnica de falsos pares (Lambert; Lambert, 1975). Nesse trabalho, 200 informantes avaliaram o provável nível de escolaridade (fundamental, médio ou superior) de ledores de estímulos sonoros com produção de sentenças em que as vogais pretônicas de algumas palavras foram produzidas ora com vogal anterior média, ora alteada.

A análise dos resultados desse experimento desenvolvida neste trabalho pretende investigar a influência, sobre as respostas dos participantes, *da estrutura da sílaba nucleada pela vogal alteada* e *da frequência de aplicação do processo no item lexical*. Portanto, somente o conjunto de respostas aos estímulos com produção [i] da vogal pretônica é assumido como VD. Já as variáveis independentes (VI) controladas são as *estruturas silábicas* 1. travada por sibilante ([i]studo, [i]scola), 2. por rótico (s[i]rviço, t[i]rceira) e 3. por nasal ([i]nsino, [i]mpresa). É controlada, também, a *frequência de ocorrência* da vogal [i] no item lexical nos *corpora* Nurc-RJ e Concordância, envolvendo itens alteados com *alta frequência* (m[i]nino, p[i]rigo) e com *baixa frequência* (acad[i]mia, v[i]stibular) nessa variedade. A relevância dos fatores estruturais mencionados é hipotetizada por Souza Guerreiro (2023), a partir da literatura prévia sobre o alteamento pretônico na fala do Rio de Janeiro (Avelheda, 2013; Souza Guerreiro, 2017). Na seção

[3] Método de coleta de dados proposto por Likert (1932) para análise de atitudes e opiniões.

6, em que discutimos a interpretação dos resultados a que chegamos, o trabalho de Souza Guerreiro (2023) é retomado com maiores detalhes.

PREPARAÇÃO DO ARQUIVO PARA IMPORTAÇÃO

Para manipulação de dados no R, utilizamos a interface RStudio (Posit Team, 2023; R Core Team, 2022). Anteriormente à importação e à leitura dos dados nesse programa, é necessário compilá-los de acordo com uma estrutura determinada, que coincide com o *input* já usado, por exemplo, pelo pacote *Rbrul*. Tal estrutura é caracterizada como um conjunto de colunas que representam cada variável considerada e cujas linhas dizem respeito às suas respectivas observações/ocorrências. A primeira linha de cada coluna pode ser dedicada à nomeação da variável. As linhas seguintes, por sua vez, registram suas ocorrências. O Quadro 1, abaixo, ilustra a estruturação mencionada.

Quadro 1 – estrutura do arquivo de dados

VD	VI 1	VI 2	VI 3	VI *n*
Observação 1	Observação 1	Observação 1	Observação 1	Observação 1
Observação 2	Observação 2	Observação 2	Observação 2	Observação 2
Observação *n*	Observação *n*	Observação *n*	Observação *n*	Observação *n*

Fonte: elaborado pelos autores

A compilação do arquivo de dados pode ser feita diretamente em editores de planilhas como *Calc* (*LibreOffice*) e *Excel* (*Microsoft Office*) ou exportada no formato desejado a partir de outros *softwares*, como *Elan*. Para leitura da planilha no R, utilizamos o formato .CSV UTF -8, embora o *software* permita, com diferentes funções, a importação de dados em outros formatos.

IMPORTAÇÃO E PREPARAÇÃO DOS DADOS NO RSTUDIO

O primeiro passo de qualquer manipulação de dados no RStudio é a definição do diretório de trabalho, isto é, a pasta do computador em que estão armazenados os arquivos com que pretendemos trabalhar. Esse processo pode ser feito manualmente, clicando em *Go to directory*, na aba inferior direita *Files*, e navegando até a pasta desejada. Após localizar e selecionar a pasta, os arquivos nela contidos devem ficar disponíveis para visualização em *Files*. Então, clicamos no ícone de engrenagem ⚙ nessa mesma aba e selecionamos *Set as working directory*[4].

Após esse processo, é necessário instalar os pacotes de funções com que lidaremos. Esse passo é necessário apenas uma vez e pode ser feito com a função *install.packages()*. Funções, em linguagem R, são códigos que executam determinado comando predeterminado. Os objetos de tais comandos são informados pelo usuário, geralmente, dentro dos parênteses (). Uma vez que pretendemos instalar pacotes específicos, no caso, *Tidyverse* e *Ordinal*, devemos informar seus nomes dentro dos parênteses, entre aspas duplas, como exemplificado em (1). Uma vez instalados, esses pacotes devem ser carregados com a função *library()*, como demonstra (2), a fim de

[4] Definido o diretório de trabalho por esse método, a linha de comando correspondente a essa ação deve aparecer na tela inferior esquerda *Console*. É recomendável que esse mesmo código seja adicionado ao *script* para garantir retomadas mais rápidas em eventuais novas sessões no RStudio.

torná-los acessíveis ao R durante a sessão atual. Esse passo deve ser repetido a cada nova sessão na ferramenta, diferentemente da instalação.

Quadro 2 – *script* de importação e preparação

```
# (1) Instalar pacotes
install.packages("tidyverse")
install.packages("ordinal")

# (2) Carregar pacotes
library(tidyverse)
library(ordinal)

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

# (4) Informar a natureza ordinal da VD e ajustar a ordem dos níveis
esc.pretonicas.ord = esc.pretonicas %>%
        mutate(VD = factor(VD,
                           levels = c("fundamental", "medio", "superior"),
                           ordered = TRUE))
```

Fonte: elaborado pelos autores a partir de Garcia (2021)

Carregados os pacotes, estamos finalmente aptos a importar o arquivo de dados com a função *read_csv2()*, parte do pacote *Tidyverse*. Com uso do operador "=", atribuímos um nome a nosso arquivo de dados, com o qual o identificaremos como um objeto na interface do RStudio. No exemplo em (3), escolhemos chamar nosso objeto de **esc.pretonicas**. O operador "=" indica que **esc.pretonicas** deve corresponder ao resultado da operação aplicada pela função *read_csv2()*, que toma como argumento o nome do arquivo **dados.csv**, armazenado em nosso diretório de trabalho.

Ao resultado de *read_csv2()* é aplicado o operador *pipe* "%>%", que, fundamentalmente, habilita um encadeamento de ações. Nesse caso, seu efeito prático é tomar o objeto carregado por *read_csv2()* e aplicar a função *mutate_if()*, que, por sua vez, promove transformações nas variáveis do arquivo de dados de acordo com os argumentos especificados: se, no arquivo, houver alguma variável do tipo *caracter*, esta deve ser convertida em uma variável do tipo *factor*[5].

Quando a linha de código em (3) é executada, o objeto **esc.pretonicas** é disponibilizado na aba superior direita *Environment*. Agora, como passo preparatório, resta apenas informar ao R que nossa VD apresenta uma relação ordinal entre seus níveis. Para isso, como exemplifica (4), criamos um novo objeto, que chamamos de **esc.pretonicas.ord**, correspondente ao resultado da aplicação da função *mutate()* a **esc.pretonicas** (lembre-se da relação de encadeamento promovida pelo *pipe*). Diferentemente de *mutate_if()*, *mutate()* não tem aplicação condicional: essa função é aplicada invariavelmente a uma coluna específica de **esc.pretonicas**: VD. Com a sintaxe da função *mutate()* apresentada em (4), estabelecemos que, no novo objeto **esc.pretonicas.ord**, a coluna VD deve ser tratada como uma variável do tipo *factor*, cujos níveis são, do menor para o maior, "fundamental", "medio" e "superior" e que essa ordenação é parte da natureza da variável (por isso, assinalamos *ordered = TRUE*).

[5] *Caracter* e *factor* são algumas classes de variáveis em linguagem R. A primeira trata as observações como ocorrências textuais. A consequência de tal tratamento é que observações idênticas são reconhecidas como distintas, ou seja, não há estabelecimento de "níveis" de variáveis por agrupamento. Já a segunda classe permite que observações idênticas sejam contadas como diferentes ocorrências do mesmo nível de uma variável categórica.

CONSTRUÇÃO DE MODELOS

Com o arquivo carregado e ajustado, avançamos para a etapa de análise. Naturalmente, toda análise inferencial deve ser precedida de uma análise exploratória exaustiva por parte do pesquisador, a fim de 1. checar a distribuição de dados em função das VI controladas e 2. derivar, pelo cotejo dos padrões distribucionais com a literatura disponível, hipóteses que fundamentem os modelos inferenciais a serem testados. Métodos para a promoção de análises exploratórias no R podem ser consultados nos manuais especializados recomendados ao final deste artigo. As descrições aqui apresentadas abrangem apenas os procedimentos de análise inferencial e pressupõem a investigação anterior da distribuição dos dados.

Quadro 3 – sintaxe básica de modelos ordinais de efeitos mistos

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

Fonte: elaborado pelos autores

O Quadro 3 apresenta a estrutura básica de um modelo ordinal construído com a função *clmm()*. A estrutura não difere muito da utilizada em funções mais populares, como *glm()*: o primeiro argumento é a VD da pesquisa – a variável cujo comportamento pretendemos explicar. Essa propriedade é codificada por "~", que representa o limite entre a VD e o(s) argumento(s) seguinte(s): as VI, possíveis fatores condicionantes da variabilidade da VD. Não há limitação nem sobre a quantidade de variáveis independentes codificadas, nem sobre sua natureza (quantitativa ou qualitativa)[6]. No Quadro 3, como lidamos com mais de uma VI, estabelecemos a intercalação entre elas com uso de "+". Esse símbolo, contudo, considera a atuação isolada de cada VI sobre a VD. Além desse cenário, é possível que duas VI atuem em conjunto, configurando uma interação. Para codificar essa possibilidade analítica no R, substituímos "+" por "*".

Na hipótese de a pesquisa incluir alguma variável aleatória – no nosso caso, o informante –, essa deve ser introduzida por "+", seguindo a notação (1|**VAR**) assumida pelo pacote, em que, à direita da barra vertical, é especificado o fator de aleatoriedade. Por fim, com o último argumento, *data* = **esc.pretonicas.ord**, especificamos, à direita de "=", o objeto do R em que estão contidos os dados a serem modelados.

COMPARAÇÃO ENTRE MODELOS

O resultado da sintaxe apresentada na última seção é um modelo em que as respostas dos avaliadores à pergunta constituinte do instrumento de pesquisa (*qual é o provável nível de escolaridade desse indivíduo?*) são explicadas pela atuação isolada das variáveis *frequência* de alteamento da vogal pretônica no item lexical e *estrutura silábica*. Após a execução do código (5), já é possível consultar no *console*, com a função *summary()*, os resultados e suas estatísticas para interpretação. Isso não significa, contudo, que a geração do modelo termine nessa etapa. O pesquisador deve, ainda, verificar a relevância da inclusão de uma variável em detrimento de outra e examinar a contribuição, para o poder explicativo do modelo, de possíveis interações hipotetizadas a partir da distribuição cruzada dos dados.

[6] Em qualquer programa probabilístico, a quantidade de dados pode ser uma limitação, tendo em vista que uma amostra muito reduzida tem o potencial de gerar resultados menos confiáveis, especialmente quando associada a um grande número de VI.

Esse tipo de análise é conduzido, no caso de modelos ordinais, por testes de estimativa por máxima verossimilhança (*likelihood ratio tests*), que podem ser realizados com a função *anova()*. Para demonstrar esse processo, primeiro, construímos um segundo modelo, com o código (6) do Quadro 4, em que passamos a considerar uma possível interação entre nossas VI. Posteriormente, com o código (7), estabelecemos a efetiva comparação entre os dois modelos.

Quadro 4 – comparação entre modelos de efeitos fixos e mistos

```
# (6) Gerar modelo 2 (efeitos mistos)
mod2 = clmm(VD ~ FREQUENCIA * ESTR.SILABICA +
                 (1|INFORMANTE), data = esc.pretonicas.ord)

# (7) Comparar modelos
anova(mod1, mod2)
```

Fonte: elaborado pelos autores a partir de Christensen (2022)

A função *anova* em (7) toma como argumentos os dois modelos construídos. Seu resultado, em todas as aplicações, deve ser semelhante ao apresentado no Quadro *5*, que reproduz o *output* do *console*.

Quadro 5 – *output* da função *anova*

```
Likelihood ratio tests of cumulative link models:

      formula:                                          link: threshold:
mod2  VD ~ FREQUENCIA * ESTR.SILABICA + (1 | INFORMANTE) logit
      flexible
mod1  VD ~ FREQUENCIA + ESTR.SILABICA + (1 | INFORMANTE) logit
      flexible

      no.par AIC    logLik  LR.sta df Pr(>Chisq)
mod2  6      515.46 -251.73
mod1  7      515.28 -250.64 2.179  1  0.1399
```

Fonte: elaborado pelos autores

A primeira linha do *output* informa o tipo de teste executado. Logo abaixo dessa informação, são retomadas as fórmulas de cada um dos modelos comparados. Por fim, temos acesso às estatísticas de qualidade de cada modelo. Em suma, a função *anova()* realiza uma avaliação das diferenças entre os elementos cotejados. Caso não haja distinção significativa entre ambos, o teste exibirá como *output* um *valor de p* superior ao nível de significância ($\alpha = 0.05$). No entanto, se houver diferença significativa, o teste retornará um *valor de p* inferior ao nível α.

O objetivo da análise inferencial é alcançar o maior poder explicativo por meio da mais simples modelagem. Como consequência disso, na comparação entre modelos, quando $p>0.05$, optamos pelo mais simples entre os comparados: o que contém menos variáveis ou, no nosso caso, o que não codifica interação entre previsores. No entanto, quando o *output* da função *anova* exibe $p<0.05$, a significância da diferença entre os modelos impede que escolhamos o mais simples. Nessa situação, o mais complexo deve ser o assumido como modelagem ótima.

INTERPRETAÇÃO DOS RESULTADOS

Antes de discutir os pormenores interpretativos do modelo selecionado na seção anterior, consideramos relevante retomar alguns detalhes e hipóteses da pesquisa de Souza Guerreiro (2023), a fim de contextualizar os dados com que lidamos e os objetivos que pretendemos atingir com a análise de regressão logística ordinal.

Os participantes da pesquisa de Souza Guerreiro (2023) ouviram nove áudios com a vogal pretônica alteada. O objetivo da autora era o de observar qual *escolaridade* os informantes tendem a atribuir ao falante que realiza o alteamento da vogal anterior. Após a audição dos estímulos, os informantes responderam a seguinte pergunta: *Se você tivesse que atribuir um grau de escolaridade a esse(a) falante, você diria que ele(a) cursou: (i) ensino fundamental I, antigo primário; (ii) ensino fundamental II, antigo ginásio; (iii) ensino médio; (iv) ensino superior*. Para a análise dos resultados, foram amalgamadas as variáveis *ensino fundamental I, antigo primário* e *ensino fundamental II, antigo ginásio*, resultando, portanto, em uma variável dependente com 3 níveis. A análise desenvolvida, desse modo, pretendeu identificar se os participantes tendem a atribuir baixa escolaridade ao falante que realiza o alteamento da vogal anterior.

Neste trabalho, com o objetivo didático de apresentar a ferramenta analítica, observamos a importância somente da *estrutura silábica* e da *frequência* para a avaliação do alteamento pretônico. Especificamente, são analisadas a avaliação dos falantes em relação ao alteamento da vogal anterior em sílaba travada por sibilante (ex: [i]scola, [i]squina) e travada por rótico (ex: s[i]rviço, s[i]rvidos).

Estudos anteriores de produção (Avelheda, 2013; Souza Guerreiro, 2017), ao analisar o contexto silábico, observaram, a partir de um *continuum*, que o alteamento da vogal anterior tende a ocorrer em estruturas silábicas travadas por sibilante (ex.: [i]scola, [i]studo) e por nasal (ex.: [i]mpregada, [i]nsino). Entretanto, as pesquisadoras verificaram que o alteamento é pouco recorrente em estrutura de sílaba livre (ex.: s[e]nhor, p[e]queno) e sílaba travada por rótico (ex.: s[e]rviço, t[e]rceira).

Logo, considerando os resultados das pesquisas de produção, criou-se a hipótese de que os informantes avaliariam: (i) positivamente o falante que produz o alteamento em sílaba travada por sibilante, pois o fenômeno é recorrente nessa estrutura silábica e (ii) negativamente o falante que realiza o alteamento em sílaba travada por rótico, pois o fenômeno tende a não ser produzido nesse contexto silábico. Tais hipóteses contemplam exatamente aspectos concernentes à *estrutura da sílaba* e à *frequência do alteamento*, justificando sua inclusão nos modelos de regressão anteriormente exemplificados.

Munidos dessas informações, o código (6) do Quadro 4, correspondente a nosso modelo ótimo, pode ser agora retomado. Para visualizar os resultados, utilizamos a função *summary()*, tomando nosso modelo como argumento, como exemplifica o código (8) do Quadro 6, cujo *output*, reproduzido no Quadro 7, são os resultados da análise de regressão.

Quadro 6 – visualização de resultados do modelo

```
# (8) Solicitar exibição dos resultados no console
summary(mod2)
```

Fonte: elaborado pelos autores

Quadro 7 – *output* parcial da função *summary()* no *console*

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
FREQUENCIAbaixa	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
fundamental | medio	-0.9449	0.2546	-3.711
medio | superior	0.9722	0.2557	3.803

Fonte: elaborado pelos autores

O *output* no Quadro 7 é originalmente antecedido de algumas informações: a identificação do tipo de regressão, a fórmula do modelo, o conjunto de dados e a especificação das variáveis aleatórias. Para preservação de espaço, contudo, optamos por não as reproduzir. A observação do Quadro 7 evidencia dois conjuntos de resultados. O primeiro, *coefficients*, apresenta as estatísticas referentes a cada nível das variáveis independentes do modelo, à exceção dos níveis de referência. Como a variável *frequência* é binária, apenas uma comparação é suficiente para dar conta da oposição entre seus níveis. A estimativa exibida, portanto, corresponde à oposição entre *baixa* e *alta frequência* quando as demais variáveis do modelo estão em seus respectivos níveis de referência. Por outro lado, a variável *estrutura silábica* é ternária. Logo, duas comparações são necessárias para abranger todas as oposições possíveis: uma que contraponha o travamento por nasal (que é o nível de referência) ao travamento por rótico e outra que contraste a referência com o travamento por sibilante.

As estimativas geradas para cada uma das oposições do modelo são medidas em *logodds*, unidade centralizada em 0. Geralmente, na regressão logística convencional, valores em *logodds* acima de 0 indicam favorecimento, e, abaixo de 0, desfavorecimento. No caso da regressão logística ordinal, porém, as estimativas devem ser lidas em função do segundo grupo de resultados, os *Threshold coefficients*. *Thresholds* indicam numericamente, em *logodds*, os limites entre os níveis da variável dependente analisada. Pelo cotejo entre os coeficientes das variáveis independentes e os de limites, portanto, é possível identificar qual dos níveis da VD tende a ser mais provavelmente atribuído de acordo com a variação dos diferentes previsores.

Vejamos, na prática, a partir dos dados analisados, o que essa relação entre coeficientes de VI e coeficientes de limites significa. Submetidas as variáveis *estrutura silábica* e *frequência* ao modelo, foram selecionadas como significativas as variáveis independentes *sílaba travada por rótico* e *sílaba travada por sibilante*.

Para a compreensão da análise, apresenta-se como é feita a leitura dos resultados. Após selecionar a rodada mais significativa, extrai-se do Quadro 7 a seguinte escala de limites (*Threshold coefficients*) medidos em *logodds* para os níveis da variável dependente *escolaridade*:

Tabela 1 – coeficientes de limites

Avaliação alteamento – Escolaridade – Coeficiente limite[7]
Ensino fundamental: < -0.95 logodds
Ensino médio: -0.95 logodds a 0.97 logodds
Ensino superior: > 0.97 logodds

Fonte: elaborado pelos autores

De acordo com os resultados acima, entende-se que: (i) se a estimativa de uma VI é <-0.95 *logodds*, há uma maior probabilidade de os informantes escolherem, pelo menos, a escolaridade de *ensino fundamental* para o falante da variante alteada; (ii) se a estimativa se encontra entre -0.95 e 0.97 *logodds*, os informantes tendem a selecionar a escolaridade *ensino médio* e (iii), se a estimativa é > 0.97 *logodds*, há uma maior probabilidade de os informantes marcarem *ensino superior*. Com

[7] Na reprodução dos resultados em *logodds*, foram consideradas apenas duas casas decimais.

base nessa orientação, passa-se à leitura dos resultados de estimativas das variáveis independentes, extraídas do Quadro 7.

Tabela 2 – reporte do modelo

Avaliação alteamento - Escolaridade			
Ensino fundamental: < -0.95 *logodds*			
Ensino médio: -0.95 a 0.97 *logodds*			
Ensino superior: > 0.97 *logodds*			
Variáveis selecionadas	**Estimativa**	**Valor de P.**	**Escolaridade mínima atribuída**
Sílaba travada por /R/	-1.94 *logodds*	$2.73e^{-06}$	Ensino Fundamental
Sílaba travada por /S/	0.57 *logodds*	0.03	Ensino Médio

Fonte: elaborado pelos autores

Ao analisar os resultados da variável *sílaba travada por rótico*, observa-se que a estimativa -1.94 *logodds* está compreendida nos resultados relativos ao *ensino fundamental* (< -0.95 logodds). Logo, de acordo com o resultado, se a vogal anterior alteada se encontra em estrutura silábica travada por rótico, como s[i]rviço, os informantes tendem a eleger escolaridade de *ensino fundamental* ao falante.

Em relação à variável *sílaba travada por sibilante*, a estimativa 0.57 *logodds* encontra-se entre os resultados referentes ao *ensino médio* (entre -0.95 e 0.97 l*ogodds*). Assim, o resultado indica que os informantes tendem a atribuir *ensino médio* ao falante que realiza o alteamento da vogal anterior em sílaba travada por sibilante, como [i]scola.

Portanto, os resultados do estudo de percepção confirmam a hipótese levantada por Souza Guerreiro (2023) de que os informantes tendem a atribuir *baixa escolaridade* ao falante que realiza o alteamento da vogal anterior travada por rótico (ex.: s[i]rviço, s[i]rvidos, t[i]rceira), possivelmente porque o fenômeno tende a ser pouco produzido nesse contexto silábico. No entanto, os informantes tendem a atribuir *ensino médio* ao falante que produz o alteamento da vogal anterior em sílaba travada por sibilante (ex.: [i]studo, [i]scola, [i]scritório). Isso indica que, no contexto em que a realização alteada é frequentemente produzida, sua realização não parece ser socialmente saliente (Avelheda, 2013; Souza Guerreiro, 2017)[8].

AVALIAÇÃO DO POTENCIAL DA FERRAMENTA

A análise inferencial de dados ordinais por meio da função *clmm()* do pacote *Ordinal* (Christensen, 2022) apresenta o potencial de contribuir para o refinamento metodológico de pesquisas de avaliação subjetiva no âmbito da Sociolinguística, uma vez que viabiliza o adequado tratamento de variáveis cujos níveis estabeleçam alguma relação hierárquica. Entre as vantagens de sua utilização, elencamos as seguintes:

1. não há limitação de níveis da variável dependente, o que representa uma importante vantagem em relação ao uso estrito da regressão logística binária. A regressão ordinal, contudo, não é necessariamente uma substituta da binária, uma vez que ambos os métodos servem a dados de naturezas diferentes. A ampliação do repertório de métodos analíticos, contudo, viabiliza o desenvolvimento de projetos com diferentes desenhos

[8] Considerando as categorias de Labov ([1966] 2001; [1972] 2008), indicador, marcador e estereótipo, Souza Guerreiro (2023) interpreta, a partir desses resultados, que o alteamento em estrutura silábica travada por /S/ apresenta característica de marcador. Já em estrutura silábica travada por rótico, o alteamento tem característica de estereótipo. Sobre isso, ler Souza Guerreiro (2023).

metodológicos que gerem diferentes tipos de dados e que respondam a diferentes perguntas de pesquisa;

2. não há limitação sobre a natureza das variáveis independentes, como há, por exemplo, no *Varbrul* e no *GoldVarb X*. Por ser ambientado no *software* R, o *Rbrul* também fornece esse benefício. Desse modo, com o *Ordinal* e o *Rbrul*, é possível incorporar variáveis contínuas, como a *idade dos indivíduos*, sem a necessidade obrigatória de discretizá-las, segmentando-as em faixas etárias. Desse modo, o analista tem maior flexibilidade no que diz respeito ao modo de tratamento de seus dados, que deve ser escolhido em função dos objetivos de pesquisa;
3. é possível construir modelos de efeitos mistos. Essa é uma vantagem compartilhada entre muitas das ferramentas ambientadas no R, incluindo *Rbrul* e *Ordinal*. O ideal é que variáveis aleatórias sejam incluídas no modelo sempre que compuserem o pacote de variação da pesquisa, a fim de confirmar se os efeitos das variáveis fixas são mantidos (Oushiro, 2022);
4. pela consideração de probabilidades cumulativas, as previsões do modelo são produzidas de maneira matematicamente adequada à natureza do dado, respeitando o ordenamento entre os níveis da variável dependente, e
5. o *output* dos resultados reflete de maneira bastante intuitiva as tendências de mudança nas previsões de acordo com a variável independente analisada. Desse modo, é possível visualizar sua distribuição probabilística na escala hierárquica estabelecida entre os níveis da variável dependente.

CONCLUSÃO

No âmbito da Sociolinguística, testes de avaliação subjetiva são realizados com diferentes objetivos e sob desenhos metodológicos variados. Consequentemente, são gerados dados de diferentes naturezas, que devem ser analisados de acordo com suas propriedades específicas. Neste artigo, discutimos uma ferramenta de análise de dados ordinais, que são um dos produtos possíveis de testes de percepção. O pacote *Ordinal*, nesse contexto, oferece recursos estatísticos apropriados para análise de variáveis ordinais, fornecendo resultados probabilísticos calculados com base na hierarquia estabelecida entre seus níveis, como observado em relação à variável *escolaridade*, cujos níveis são organizados em uma ordem determinada: *ensino fundamental* < *ensino médio* < *ensino superior*.

A utilização desse método permitiu visualizar, a partir dos dados de Souza Guerreiro (2023), que os informantes tendem a avaliar o falante da variante alteada [i] em sílaba travada por rótico (*s[i]rviço*) como uma pessoa com nível fundamental de ensino. Há, portanto, certa rotulação social desse falante como um indivíduo de pouca escolaridade. Já em relação à sílaba travada por sibilante (*[i]scola*), os participantes tendem a atribuir escolaridade de nível médio. Esse resultado pode ser considerado adequado porque o modelo empregado considera a hierarquia estabelecida entre os níveis da VD no cálculo de suas previsões probabilísticas.

Se, para os dados analisados, a hierarquia é um fator determinante da ferramenta adotada, em outros casos, pode não o ser. Por essa razão, o pesquisador deve estar atento à natureza do dado gerado, uma vez que são suas características que determinam o método mais adequado de análise. Diante disso, é desejável que diferentes ferramentas analíticas sejam conjugadas no repertório do analista para que consiga responder suas questões de pesquisa, sem que isso seja um fator limitante dos desenhos metodológicos empregados.

Esperamos que este trabalho contribua com a disseminação de um método inferencial possivelmente aplicável às análises de estudos de percepção. O tema, contudo, não se encerra neste artigo. Por essa razão, outras leituras são recomendadas. Christensen (2022), autor do pacote

estatístico *Ordinal*, documenta exaustivamente as funções que integram a ferramenta. A função *clm()*, desse mesmo pacote, é explorada por Garcia (2021), que descreve seu modo de utilização. Essa função é análoga à *clmm()*, mas é adequada apenas para análise de variáveis fixas. A construção de modelos mistos com o pacote *Ordinal*, como aqui fizemos, é exemplificada em detalhes por Barlaz (2022) em seu *site* pessoal. Baayen (2008), por sua vez, descreve, além da regressão ordinal, métodos de regressão linear e logística. O autor dedica, ainda, um capítulo especial à discussão de modelos mistos. Levshina (2015) também aborda variados métodos de análises uni e multivariadas. Já em língua portuguesa, há os trabalhos de Oushiro (2015, 2022) e de Gries (2019), além dos cursos *Análise Quantitativa de Dados em Linguística* (Lima Jr., 2021) e *Estatística para as Ciências da Linguagem* (Godoy, 2021), disponíveis em *playlists* públicas do *YouTube*. Há, ainda, na plataforma EAD da Associação Brasileira de Linguística (Abralin), os cursos *Modelos de Regressão para Linguistas* (Lima Jr.; Garcia; Angele, 2020) e *Introdução à Estatística com R* (Oushiro, 2021).

A interface entre Linguística e Estatística, portanto, tende a colaborar com a melhoria e com a adequação de processos analíticos. Novos métodos de análise e atualizações dos já existentes surgem sempre no campo estatístico. Por isso, é importante que seja constantemente buscada a interação entre as duas áreas, a fim de que seja avaliada a relevância da incorporação de novas metodologias em nossas análises. Isso não significa, contudo, que a Estatística seja o fim do trabalho, mas um *meio*: um método a contribuir com o tratamento e com a interpretação de dados, a partir dos quais são derivadas as reflexões sobre o conhecimento linguístico.

REFERÊNCIAS

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.