Análise quantitativa no estudo da variação linguística: noções de estatística e análise comparativa entre Varbrul e SPSS

Quantitative analysis in the study of language variation: notions of statistical and comparative analysis Varbrul and SPSS

Alan Jardel de Oliveira
Universidade Federal de Minas Gerais - UFMG

Resumo

Este artigo apresenta um estudo do modelo estatístico utilizado nos estudos variacionistas e uma análise das especificidades do *software* Varbrul em relação aos métodos convencionais de estimação de parâmetros. Além disso, apresenta os aspectos teóricos estatísticos mais relevantes que contribuem para a análise da variação linguística e uma análise comparativa entre os resultados do Varbrul e do SPSS.

Palavras-chave

Sociolinguística variacionista, Métodos quantitativos em linguística, Softwares Varbrul e SPSS

Abstract

This article presents a study of a statistical model used in the variational studies and an analysis of specifity of Varbrul compared to conventional methods of parameter estimation. It also presents the most relevant theoretical aspects of statistics that contribute to the analysis of language variation and a comparison between the results of Varbrul and SPSS.

Keywords

Variationist sociolinguistics, Quantitative methods in linguistics, Varbrul and SPSS softwares

1. Introdução

A análise quantitativa está na base dos estudos sociolinguísticos variacionistas. Desde os primeiros estudos em variação linguística (FISHER (1974) [1958], LABOV (1963) e LABOV (1964)) tem-se adotado modelos quantitativos para dar suporte à necessidade de se estudar a linguagem em uso e a variação linguística na forma como ela ocorre na fala de uma comunidade linguística. A identificação da variabilidade ocorrida nas línguas como um fenômeno dependente de determinadas variáveis, e não como uma "variação livre" como propunham os estruturalistas, foi possível a partir do desenvolvimento de técnicas eficazes de análise de tal variabilidade. A homogeneidade do sistema linguístico e a ocorrência da "variação livre" na perspectivas estruturalista deram lugar, nos estudos variacionistas, à heterogeneidade, passível de observação e de quantificação, e a uma variabilidade condicionada por fatores sociais e por fatores linguísticos.

A partir de uma pesquisa bibliográfica na área da sociolinguística variacionista, é possível constatar que o aprofundamento no estudo dos métodos estatísticos utilizados no estudo da variação linguística não tem recebido, na maioria das vezes, a atenção merecida. No entanto, tal aprofundamento é de fundamental importância, na medida em que grande parte dos estudos variacionistas utiliza metodologias de análise quantitativa.

O *software* Varbrul, criado por David Sankoff em 1972, e aprimorado nos anos seguintes,[1] permitiu aos linguistas interessados no estudo da variação o acesso a métodos de análise estatística. Entretanto, a facilidade no acesso e na interpretação dos resultados gerados pelo Varbrul gerou também a possibilidade de utilização de métodos estatísticos sem que se tivesse necessariamente um conhecimento estatístico mais geral ou mesmo um conhecimento mais específico dos procedimentos internos utilizados pelo próprio Varbrul. Guy (1988, p. 25), num texto intitulado *Varbrul: análise avançada*, afirma que os métodos e problemas associados à análise quantitativa, entre linguistas, "(...) costumam passar de boca

em boca, uma versão acadêmica da tradição pré-letrada da história oral". O autor afirma que escreve "não como um desbravador ou inovador, mas como um mero compilador das histórias contadas por aqueles que sabem". As ideias expressas por Guy revelam uma realidade nos estudos variacionistas. Uma busca por referências bibliográficas mais aprofundadas do modelo estatístico utilizado pelo Varbrul e, consequentemente, do modelo mais utilizado pela sociolinguística, indica que todo o desenvolvimento teórico do modelo estatístico implantado no Varbrul parece concentrar-se em uma pequena quantidade de textos. Além disso, o Varbrul parece carregar consigo o mito de que somente ele é capaz de analisar com eficiência os dados coletados nos estudos variacionistas e que o método estatístico e os procedimentos realizados por ele são exclusivos e específicos para a análise linguística. Sobre esse aspecto, Guy e Zilles (2007, p. 106) afirmam que "o Varbrul tem certas vantagens que fazem dele uma boa opção para o sociolinguista. Em primeiro lugar, ele é dedicado à estruturação dos dados que encontramos na linguagem natural". Esse artigo objetiva demonstrar que o Varbrul utiliza um modelo estatístico que é amplamente difundido e utilizado nas mais diversas áreas do conhecimento e está disponível em outros *softwares* estatísticos, entretanto, ele diferencia-se por utilizar métodos pouco convencionais de codificação[2] das variáveis independentes.

Limitarmos o estudo da estatística ao estudo dos procedimentos do Varbrul e de suas referências bibliográficas leva-nos a uma limitação no desenvolvimento metodológico do estudo da variação linguística, na medida em que a compreensão mais aprofundada de outros modelos e do próprio modelo de regressão logística pode possibilitar análises mais amplas dos fenômenos linguísticos. Neste artigo será apresentado um aprofundamento do modelo estatístico utilizado nos estudos variacionistas, bem como uma interpretação mais clara das especificidades do Varbrul em relação aos métodos convencionais de estimação de parâmetros. Além disso, serão apresentados os aspectos teóricos mais relevantes da estatística que contribuem para a análise da variação linguística e uma análise comparativa entre o Varbrul e o SPSS. No caso de um interesse ainda mais aprofundado, podem-se consultar os livros de Pagano e Gauvreau (2004), Hosmer e Lemeshow (2000), Kleinbaum (1994), Johnson (2004), Paolillo (2002), entre outros.

Para a análise comparativa serão utilizados os *softwares* GoldVarbX (de agora em diante *Varbrul*) e o SPSS v.13.0 (de agora em diante *SPSS*). O banco de dados analisado nos exemplos foi criado a partir de dados reais de fala,

coletados na cidade de Itaúna/MG. Análises dos resultados referentes a tais dados podem ser encontradas em Oliveira (2006), Viegas e Oliveira (2008) e Viegas e Oliveira (2009). O fenômeno estudado em Itaúna foi a variação na sílaba final átona /l/V (lateral alveolar seguida de vogal). No estudo foram identificadas as seguintes variantes:

1. Realização plena da sílaba /l/V. Ex.: "quando não era [ˈeli], era o padre (...)" (ele), LM40.
2. Apagamento da vogal na sílaba /l/V. Ex.: "minha mãe nunca foi na [isˈkɔl] por minha causa." (escola), TH18.
3. Apagamento da vogal na sílaba /l/V e velarização de /l/. Ex.: "eu acho até que a gente era mais [trãˈkwiɬ]." (tranquilo), RH17.
4. Apagamento da sílaba /l/V. Ex.: "aí foi quando em oitenta e dois que teve [aˈkɛ] virada". (aquela), EM39.

Foram consideradas as seguintes variáveis independentes:

1. Gênero: masculino e feminino.
2. Faixa etária: jovem e adulto.
3. Contexto seguinte: consoante, vogal e pausa (essa variável refere-se ao som inicial da palavra imediatamente posterior à palavra em análise). Exemplos: ele caiu (consoante), ele entrou (vogal), falei com ele (pausa).
4. Contexto anterior: vogal alta, vogal baixa e vogal média (essa variável refere-se à vogal imediatamente posterior à sílaba /l/V). Exemplos: bula (alta), bala (baixa), bela (média).
5. Classe da palavra: nome, pronome e verbo (essa variável refere-se à classe da palavra em análise). Exemplos: janela (nome), aquela (pronome), fala (verbo).
6. Classe da palavra seguinte: auxiliar, não auxiliar, nome e pausa (essa variável refere-se à classe da palavra imediatamente seguinte à palavra em análise). Exemplos: ela está cantando (auxiliar), ela canta (não auxiliar), janela grande (nome), falei com ele (pausa).
7. Vogal na variável: [u], [i], [a] (essa variável refere-se à altura da vogal na sílaba /l/V). Exemplos: aquilo (u), aquele (i), aquela (a).

8. Tonicidade seguinte: átona, tônica, pausa (essa variável refere-se à tonicidade da sílaba imediatamente seguinte à palavra em análise). Exemplos: ele cantou (átona), ele foi (tônica), falei com ele (pausa).
9. Tipo de informação no turno: nova, dada (essa variável refere-se à repetição ou não de uma mesma palavra num turno (sem falas intermitentes do entrevistador). A primeira ocorrência da palavra no turno é codificada como *nova*, as demais ocorrências são codificadas como *dada*).
10. Presença de /S/: ausente, presente. (essa variável refere-se à presença ou ausência de /S/ na sílaba /l/V. Exemplos: ele (ausente), eles (presente).

2. A escolha do modelo estatístico

A seleção de um modelo[3] estatístico a ser utilizado se dá, primeiramente, a partir de quais perguntas se quer responder. Nos estudos em sociolinguística variacionista, relacionados à variação sonora, por exemplo, tem-se que um som ora é produzido de uma forma, ora é produzido de outra forma, em uma mesma palavra. Nesse caso, a pergunta é: o que poderia estar influenciando os indivíduos de uma mesma comunidade a falarem uma mesma palavra ora de uma forma, ora de outra? Os modelos estatísticos que permitem responder a essa pergunta, ou seja, que permitem que se possa explicar a variabilidade de um fenômeno em relação a um conjunto de fatores, são chamados de modelos de regressão. Nos modelos de regressão temos sempre uma variável, chamada variável dependente ou variável resposta, e uma ou mais variáveis explicativas, chamadas covariáveis ou variáveis independentes, que poderão ajudar a explicar a variabilidade na variável dependente.

A seleção do modelo estatístico a ser utilizado também depende do tipo de variável dependente do estudo. Se a variável dependente for contínua, por exemplo, poderíamos optar por um modelo de regressão linear. Estudos utilizando tal modelo podem ser encontrados em Labov (1994) e Labov (2001). Esse modelo poderia ser utilizado caso a variável dependente fosse, por exemplo, a duração de uma vogal, no qual teríamos observações localizadas em uma faixa contínua. Por outro lado, se a variável dependente for categórica (0 ou 1), poderíamos utilizar o modelo de regressão logística. Esse modelo poderia ser utilizado caso a variável dependente fosse composta de somente duas

possibilidades, como a presença ou a ausência da concordância verbal. Se a variável dependente fosse categórica e apresentasse mais de duas possibilidades, poderíamos utilizar o modelo multinomial. Tal modelo poderia ser utilizado, por exemplo, no estudo do pronome você, caso fossem consideradas as realizações de mais de duas variantes, consideradas por hipótese como categóricas, como você, ocê e cê.

Na maior parte dos estudos variacionistas tem-se utilizado o modelo de regressão logística, já que é esse o modelo implementado no *Varbrul*. Esse modelo é utilizado quando se quer investigar, dado um conjunto de possíveis variáveis independentes, quais delas estão de fato associadas a uma variável dependente binária (composta por duas variantes).

Tomemos o conjunto de dados de fala no qual encontramos duas possibilidades para os itens lexicais terminados em sílaba átona formada por /l/V, dadas por a) /l/V e b) Ø. Assim, itens como *aquele* e *tranquilo* poderiam ocorrer também como *aquê* e *tranqui*. Nesse caso, temos uma variável dependente binária /l/V e Ø.

Suponhamos ter a hipótese de que o gênero influencia a produção de uma ou outra forma. Assim, temos uma variável independente *gênero*, composta pelos fatores[4] *masculino* e *feminino*.

A variável dependente apresenta duas possibilidades, ou ocorre /l/V ou ocorre Ø. Podemos codificá-la então como 0 ou 1. Nesse caso, determina-se que a variante codificada como 0 seja a variante /l/V e a variante codificada como 1 seja a variante Ø, já que o objetivo é analisar o fenômeno de apagamento da sílaba final /l/V. Em estatística, normalmente a variante codificada como 1 recebe o nome de *sucesso*, em oposição à denominação *fracasso* para a variante codificada como 0.

Se utilizarmos o modelo de regressão logística para a análise da variável /l/V (0) e Ø (1), tendo como variável independente o gênero (masculino ou feminino), poderemos determinar a influência do gênero na probabilidade de sucesso da variável dependente, ou seja, na probabilidade de utilização da variante Ø. Além disso, o modelo permite que seja analisado simultaneamente o efeito de múltiplas variáveis independentes. Nas seções seguintes, o modelo de regressão logística será aprofundado.

3. Noções preliminares

Nesta seção serão apresentadas algumas noções estatísticas preliminares para que posteriormente se apresente o modelo de regressão logística e a análise comparativa entre os *softwares* Varbrul e SPSS.

3.1. Hipótese nula, nível de significância e p-valor

Uma hipótese levantada para explicar estatisticamente algum fenômeno vem associada a uma segunda hipótese que nega a primeira. Tais hipóteses são chamadas, respectivamente, de hipótese alternativa e hipótese nula.

Nos modelos de regressão, por exemplo, um teste estatístico poderia propor testar a hipótese nula de não haver efeito na variável dependente associado a uma variável independente. Assim, a suposição de que a variabilidade na sílaba final átona /l/V possa ser explicada pelo gênero dos falantes vem acompanhada da hipótese nula que sugere que o gênero dos falantes não exerce influência estatisticamente significativa sobre tal variabilidade. A hipótese efetivamente testada é a hipótese nula.

A probabilidade máxima aceitável de rejeitarmos a hipótese nula, quando ela é de fato verdadeira, é chamada de *nível de significância*. No exemplo acima, o nível de significância seria a probabilidade máxima de cometermos um erro ao aceitarmos que o gênero do falante interfere na variabilidade, quando na realidade ele não interfere. O *nível de significância* é um valor arbitrário, definido segundo critérios do pesquisador. Convencionalmente, na sociolinguística variacionista, assim como em outras áreas, utiliza-se um *nível de significância* de 0,05.

Um teste estatístico pode fornecer a probabilidade de o efeito observado ser proveniente do acaso. Tal probabilidade é chamada de *p-valor*. Uma maneira de conduzir um teste estatístico é o de rejeitar a hipótese nula quando o *p-valor* é menor que o nível de significância. Nesse caso, se encontramos um *p-valor* de 0,02 em um teste estatístico, podemos afirmar que a hipótese nula foi rejeitada, já que o *p-valor* foi menor do que o nível de significância de 0,05. Isso indica que os resultados obtidos são estatisticamente significativos.

3.2. Probabilidade, chance e razão de chances

A probabilidade pode ser definida como uma medida numérica da possibilidade de ocorrência de um evento qualquer em uma população. Na

prática, ela pode ser obtida pela razão entre o número de ocorrências de um evento e o número total de ocorrências da amostra. Vejamos a tabela de contingência[5] abaixo, obtida a partir da classificação de uma amostra de 2.280 observações segundo o gênero e a variável dependente /l/V ~ Ø:

TABELA 1
Tabela de Contingência para o gênero na variável /l/V ~ Ø

	Masculino	**Feminino**	**Total**
/l/V	356	715	1071
Ø	588	621	1209
Total	944	1336	2280

Com base nos dados acima, podemos calcular a probabilidade de um indivíduo utilizar a variante Ø, bem como a probabilidade associada aos gêneros:

$$\hat{p}_{total} = \frac{1209}{2280} = 0,53$$

$$\hat{p}_{masc} = \frac{588}{944} = 0,62$$

$$\hat{p}_{fem} = \frac{621}{1336} = 0,46$$

sendo $\hat{p}_{total}$ a probabilidade total de ocorrência de Ø, $\hat{p}_{masc}$ a probabilidade de Ø dado que o indivíduo pertence ao gênero masculino e $\hat{p}_{fem}$ a probabilidade de Ø dado que o indivíduo pertence ao gênero feminino.

Como a probabilidade é sempre um número entre 0 e 1, se temos a probabilidade de ocorrer um evento (p), a probabilidade de que tal evento não ocorra será 1–p. Assim, a probabilidade de ocorrer Ø em um indivíduo do gênero masculino é 0,62 e a probabilidade de não ocorrer Ø é 1–0,62=0,38.

Outra medida importante é a *chance*, ou *odds*, definida como a razão entre a probabilidade de que um evento ocorra e a probabilidade de que ele não ocorra. Assim, a *chance* para as probabilidades acima é dada por:

$$odds_{total} = \frac{0,53}{1 - 0,53} = 1,13$$

$$odds_{masc} = \frac{0,62}{1 - 0,62} = 1,65$$

$$odds_{fem} = \frac{0,46}{1 - 0,46} = 0,87$$

A interpretação de um resultado em termos de chance é feita da seguinte forma: a probabilidade de ocorrer Ø entre os homens é 1,65 vezes a probabilidade de não ocorrer Ø, ou seja, de ocorrer /l/V; e a probabilidade de ocorrer Ø entre as mulheres é 0,87 vezes a probabilidade de ocorrer /l/V. Com base nesses resultados, podemos inferir que a chance de ocorrer a variante Ø entre os homens é quase duas vezes a chance de ocorrer entre as mulheres (1,65 para 0,87). A comparação entre as chances de ocorrência de um evento entre fatores de uma variável é chamada de *razão de chances* e é bastante utilizada nos estudos que utilizam modelos de regressão logística. A razão de chances, ou *odds ratio*, fornece uma medida do grau de associação entre fatores de uma variável independente. No exemplo acima, a razão de chances entre o gênero masculino e o gênero feminino é obtida através da razão entre as *chances* dos gêneros,

$$OR = \frac{odds_{masc}}{odds_{fem}} = \frac{1,65}{0,87} = 1,9$$

A razão de chances de 1,9 indica que a chance de *sucesso* (nesse caso, a variante Ø) para o gênero masculino é 1,9 vezes a chance no gênero feminino. Isso indica que a chance de um homem, na cidade de Itaúna, utilizar a variante Ø é quase duas vezes a chance de uma mulher utilizar a mesma variante.

4. O modelo de regressão logística

O modelo de regressão logística é utilizado quando temos uma variável dependente binária, nos casos dos exemplos na seção anterior, /l/V ou Ø, e variáveis independentes que podem contribuir para explicarmos a variabilidade

na variável dependente. Matematicamente, o modelo de regressão logística, considerando *n* variáveis independentes, é definido pela equação:

$$log \frac{p}{(1-p)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Na equação acima,

- a função $log \frac{p}{(1-p)}$, ou simplesmente $logit(p)$, é chamada de função de ligação;
- p é a probabilidade de *sucesso*;
- α é a constante que expressa o valor do $logit(p)$ quando todas as variáveis independentes são iguais a 0;
- β (variando de 1 a n) é um coeficiente que expressa o efeito das variáveis independentes x (variando de 1 a n) na função de ligação, quando a variável x aumenta uma unidade.

Os valores de α e dos β's são estimados pelo método da máxima verossimilhança e obtidos a partir de um algoritmo numérico computacional.[6] Esse método encontra, entre todos os valores possíveis, os valores de α e dos β 's que sejam mais prováveis de terem gerado os dados observados.

Tomando, por exemplo, como variável dependente /l/V (codificada como 0) e Ø (codificada como 1) e como variável independente o gênero – feminino (0) e masculino (1) – a equação da regressão logística estimada será definida por:

$$log \frac{\hat{p}}{(1-\hat{p})} = \hat{\alpha} + \hat{\beta}_1 \text{ gênero}_1$$

onde $\hat{p}$ é a estimativa da probabilidade de ocorrência de variante Ø, $\hat{\alpha}$ é a constante e $\hat{\beta}_1$ o efeito da variável *gênero* no $logit(\hat{p})$, quando ela aumenta uma unidade, ou seja, passa de 0 (feminino) para 1 (masculino).

Rodando no SPSS a regressão logística, tendo como variável dependente /l/V (0) e Ø (1) e como variável independente o gênero, teremos o seguinte valor para β:

$$\hat{\beta}_1 = 0{,}643$$

Portanto, temos um aumento de 0,643 no $logit(\hat{p})$, quando passamos do gênero feminino (0) para o gênero masculino (1).

Utilizando-se propriedades do logaritmo e tomando-se a exponencial do coeficiente β, obtemos uma interpretacão da associação em termos de razão de chances:

$$OR = exp(\beta)$$

Assim, dado que $\hat{\beta}$ para a variável gênero é 0,643, a razão de chances entre os fatores da variável gênero será:

$$\hat{OR} = exp\ (0,643) = 1,9$$

O valor 1,9 para a razão de chances corresponde ao mesmo valor encontrado na seção anterior. Entretanto, o modelo exemplificado aqui contém somente uma variável independente. Em um modelo multivariado, a estimativa do efeito de uma variável altera-se com a inserção de outras variáveis no modelo, o que não permite que a razão de chances seja obtida das chances calculadas por meio de uma tabela de contingência, como na seção 3.2. Nesse caso, temos uma razão de chances que leva em consideração o efeito das demais variáveis independentes.

Seleção das variáveis independentes

Após definirmos o conjunto de variáveis independentes a serem incluídas no modelo de regressão logística, precisamos identificar as variáveis mais importantes para explicar a probabilidade de sucesso. Hosmer e Lemeshow (2000) afirmam que, tradicionalmente, a construção de um modelo estatístico implica a busca do modelo mais parcimonioso para a explicação dos dados.[7] Assim, interessa-nos, entre todas as variáveis independentes consideradas, identificar o conjunto de variáveis que melhor contribui para explicarmos a variabilidade.

O processo de seleção de variáveis pode ser feito de maneiras diversas. Os métodos utilizados no *Varbrul*, chamados *step-up* e *step-down*, são conhecidos em estatística como métodos *stepwise*, respectivamente como *forward* e *backward*, e são encontrados nos *softwares* estatísticos convencionais. No SPSS, eles recebem o nome de *foward-lr* e *backward-lr*.

Os métodos *stepwise* permitem que a seleção das variáveis independentes mais importantes seja feita de forma automática, segundo critérios predefinidos. Sobre os métodos *stepwise*, Hosmer e Lemeshow (2000) afirmam:

> Qualquer procedimento stepwise para a seleção ou exclusão de variáveis de um modelo baseia-se em um algoritmo estatístico que verifica a "importância" de variáveis, e que inclui ou exclui com base em uma regra de decisão fixa. A "importância" de uma variável é definida em termos de uma medida da significância estatística do coeficiente para a variável.[8] (HOSMER; LEMESHOW, 2000, p. 16).

O critério de seleção utilizado nesses métodos, em ambos os *softwares*, é a *razão da máxima verossimilhança*. No método *forward* (*step-up)*, os programas inserem, passo a passo, cada uma das variáveis independentes, uma a uma, e comparam o valor do *logaritmo da verossimilhança*[9] ou *log-likelihoods* dos modelos com e sem a variável independente adicionada.

No passo 1, obtém-se o valor do *log-likelihood* sem a inclusão de nenhum parâmetro associado às variáveis independentes, ou seja, considera-se somente o parâmetro $\hat{\alpha}$. No passo seguinte, testam-se todas as variáveis inseridas no modelo uma a uma e seleciona-se somente aquela que gera o maior valor no *teste da razão da máxima verossimilhança*.

O procedimento se repete nos passos seguintes. As variáveis selecionadas em cada passo vão sendo mantidas no modelo utilizado para seleção da próxima variável. A seleção das variáveis independentes é interrompida quando a diferença entre o modelo sem a variável e o modelo com a variável não apresenta significância estatística (p-valor < 0,05).[10]

No método *backward* (*step-down*) selecionam-se as variáveis mais importantes tendo como referência um modelo em que todas as variáveis são incluídas. Em cada passo seguinte, testa-se cada uma das variáveis e retira-se aquela que apresenta o menor valor no *teste da razão da máxima verossimilhança*, ou seja, que apresenta o maior *p-valor*. A retirada de variáveis é interrompida quando o teste apresenta resultado estatisticamente significativo.

Os métodos *stepwise* facilitam bastante o trabalho do pesquisador, entretanto a seleção baseia-se exclusivamente em critérios estatísticos.

No SPSS, o método padrão de inserção de variáveis no modelo é o método manual, denominado *enter*. No *Varbrul*, esse procedimento é chamado de *one-level*. O método *enter* ou *one-level* permite que a entrada das variáveis independentes seja feita de forma manual, segundo critérios do pesquisador. Com base na comparação das saídas de modelos contendo conjuntos diferentes de variáveis independentes, utilizando o *teste da razão da máxima*

verossimilhança, pode-se determinar o melhor conjunto de variáveis para explicar a probabilidade de sucesso. Esse método é mais interessante, pois pode partir de decisões estatísticas e linguísticas, mas exige do pesquisador um maior domínio dos métodos de comparação de modelos para que o melhor modelo possa ser eficientemente selecionado.

5. Comparando o Varbrul e o SPSS

Como veremos nessa seção, os efeitos estimados das variáveis independentes apresentados pelo Varbrul diferem-se dos efeitos apresentados em uma saída padrão do SPSS. Como será mostrado, o resultado gerado pelo Varbrul diferencia-se pela maneira como são codificados os fatores que compõem as variáveis independentes. Essa diferença gera efeitos diferenciados para as variáveis. No SPSS, assim como em outros *softwares* estatísticos, podemos definir o parâmetro de codificação dos fatores de maneira semelhante à codificação feita internamente pelo Varbrul e, assim, obtermos resultados semelhantes.

5.1. Diferentes formas de codificação de fatores

5.1.1. Fator de referência

Numa regressão logística convencional, utiliza-se uma codificação de fatores das variáveis independentes em que se determina um fator de referência. Tal codificação é amplamente difundida e utilizada nas mais diversas áreas do conhecimento. No SPSS, assim como em outros pacotes estatísticos, a codificação com fator de referência é dada como padrão. Nela, os efeitos dos demais fatores de uma variável independente e, consequentemente, a razão de chances, estarão em relação ao fator de referência. No exemplo mostrado na seção 3.2, o fator de referência[11] para a variável gênero foi o gênero *feminino*. Assim, a razão de chances obtida (1,9) refere-se à chance do gênero *masculino* (1,9) em relação ao gênero *feminino* (1).

A codificação dos fatores é feita pelo *software* por meio de uma tabela, chamada *matriz de desenho*. Quando temos somente dois fatores, a *matriz de desenho*, tendo o primeiro fator como referência, é:

TABELA 2
Matriz de desenho para a variável gênero

Gênero	Código
Feminino	0
Masculino	1

No caso de termos uma variável com três fatores, a matriz de desenho, tendo o primeiro fator como referência, é:

TABELA 3
Matriz de desenho para a variável contexto seguinte

Contexto Seguinte	Código do parâmetro (1)	Código do parâmetro (2)
Consoante	0	0
Vogal	1	0
Pausa	0	1

Nesse caso, a equação do modelo de regressão logística é dado por

$$log\ \frac{\hat{p}}{(1-\hat{p})} = \hat{\alpha} + \hat{\beta}_1\ vogal + \hat{\beta}_2\ pausa$$

Os efeitos estimados, obtidos por meio de uma análise de regressão logística no SPSS tendo como variável dependente /l/V e Ø e como variável independente os contextos seguintes *consoante*, *vogal* e *pausa*, são:

$$\hat{\alpha} = 0{,}546$$
$$\hat{\beta}_1 = \text{-}1{,}034$$
$$\hat{\beta}_2 = \text{-}2{,}187$$

A razão de chances entre os fatores, como mostrado na seção 3, é o exponencial do $\hat{\beta}$. Assim, teremos as seguintes razões de chances para a variável *contexto seguinte*, tendo o fator *consoante* como fator de referência:

$$OR_{vogal} = exp(\text{-}1{,}034) = 0{,}36$$
$$OR_{pausa} = exp(\text{-}2{,}187) = 0{,}11$$

Logo, a chance de um indivíduo utilizar a variante Ø antes de vogal é 0,36 a chance de utilizá-la antes de consoante. Da mesma forma, a chance de um indivíduo

utilizar a variante Ø antes de pausa é 0,11 a chance de utilizá-la antes de consoante. As relações entre as OR das categorias não definidas como referência também podem ser feitas. Dessa forma, a chance de um indivíduo utilizar a variante Ø antes de vogal é 3,27 (0,36/0,11) a chance de utilizá-la antes de pausa.

5.1.2. Desvio da média

O modelo de regressão logística implementado no *Varbrul* é idêntico ao modelo implementado em outros pacotes estatísticos. Entretanto, o *Varbrul* difere-se pela maneira como os fatores das variáveis independentes são codificados. Em vez de tomar um fator como referência e, a partir dele, estimar o efeito dos demais fatores, o *Varbrul* utiliza uma codificação conhecida como *desvio da média*. No método *desvio da média*, a variável *gênero* é codificada como (-1) *feminino* e (1) *masculino*, em vez de 0 e 1, respectivamente, como no método *fator de referência*. A diferença na codificação gera diferentes efeitos estimados, como pode ser visto a seguir.

Tomando o conjunto de dados com variável dependente /l/V e Ø e variável independente o contexto seguinte *consoante, vogal* e *pausa*, temos a seguinte tabela de contingência:

TABELA 4
Tabela de contingência para o contexto seguinte na variável dependente /l/V ~ Ø

	Consoante	**Vogal**	**Pausa**	**Total**
/l/V	603	184	284	1071
Ø	1041	113	55	1209
Total	1644	297	339	2280

Por meio da tabela 4, podemos obter a chance de cada fator. O método *desvio da média* utiliza o logaritmo (ln) da chance dos fatores ($\hat{g}$), dado por:

$$\hat{g}_{cons} = ln(\frac{1041}{603}) = 0,546$$

$$\hat{g}_{vog} = ln(\frac{113}{184}) = 0,488$$

$$\hat{g}_{pausa} = ln(\frac{55}{284}) = 1,642$$

Com base nesses valores, podemos calcular também um valor médio ($\hat{g}_{media}$) dos fatores:

$$\hat{g}_{media} = \frac{\hat{g}_{cons} + \hat{g}_{vog} + \hat{g}_{pausa}}{3} = 0,\ 528$$

Os valores dos efeitos estimados ($\hat{\beta}$) a serem incluídos na equação da regressão logística, no método *desvio da média*, são dados pelos $\hat{g}$'s das categorias subtraídos da média dos $\hat{g}$'s, ou seja,

$$\hat{\beta}_1 = \hat{g}_{cons} - \hat{g}_{media} = (0,546) - (-0,548) = 1,074$$

$$\hat{\beta}_2 = \hat{g}_{vog} - \hat{g}_{media} = (-0,488) - (-0,528) = 0,040$$

$$\hat{\beta}_3 = \hat{g}_{vogal} - \hat{g}_{media} = (-1,642) - (-0,528) = -1,114$$

Com base nos *betas* calculados acima, podemos calcular uma razão de chances em relação à média (OR') para cada categoria:

OR'_{cons}=exp(1,074)=2,926

OR'_{vog}=exp(0,040)=1,041

OR'_{pausa}=exp(-1,114)=0,328

A razão de chances proveniente do método *desvio da média* não é a mesma da razão de chances calculada pelo método *fator de referência*. Hosmer e Lemeshow (2000, p. 60) afirmam que tal *razão de chances* é uma razão de chances do fator em relação à *média geométrica* das chances de todos os fatores da variável independente.[12]

Diante dos resultados apresentados anteriormente, como obter o *factor weights* ou *peso relativo*, fornecido na saída do *Varbrul*? De acordo com Morrison (2005), o *peso relativo* é dado por *OR'/(1+OR')*, ou seja, o *peso relativo* é uma medida calculada a partir da razão de chances, calculada pelo método *desvio da média*. Como a razão de chances é um número entre 0 e infinito, o *peso relativo* será sempre um número entre 0 e 1. Caso a OR' seja

igual a 1, teremos que o *peso relativo* será igual a 0,50. Assim, os *pesos relativos* para as categorias da variável *contexto seguinte* são:

$$PR_{cons} = \frac{2,926}{1+2,926} = 0,745$$

$$PR_{vog} = \frac{1,041}{1+1,041} = 0,510$$

$$PR_{pausa} = \frac{0,328}{1+0,328} = 0,247$$

O valor do *input* corresponde ao *peso relativo* da média das *chances*, dado por:

$$input = \frac{exp\ (\hat{g}_{media})}{1+exp\ (\hat{g}_{media})} = \frac{exp\ (-0,528)}{1+exp\ (-0,528)} = 0,371$$

Os resultados acima podem ser comprovados pelos resultados obtidos por uma rodada com *fatores centralizados* no *Varbrul*:

Run # 2, 3 cells:
Convergence at Iteration 5
Input 0.371
Group # 1 — C: 0.745, V: 0.510, P: 0.247
Log likelihood = -1428.074 Significance = 0.000

De acordo com Rand e Sankoff (1990), a opção *fatores centralizados*[13] no *Varbrul* considera que os fatores em um grupo possuem pesos iguais; se não selecionarmos essa opção, cada fator recebe um peso de acordo com a sua ocorrência relativa no total de ocorrências no grupo.

Se multiplicarmos os valores das frequências relativas encontrados na tabela abaixo, pelos valores encontrados para os $\hat{g}$'s dos fatores, temos um valor médio dos $\hat{g}$'s considerando a freqüência relativa das ocorrências em cada fator particular. Assim, temos:

TABELA 5
Tabela de contingência para variável dependente /l/V ~ Ø com frequência relativa dos fatores da variável *contexto seguinte*

	Consoante	**Vogal**	Pausa	Total
/l/V	603	184	284	1071
Ø	1041	113	55	1209
Total	1644	297	339	2280
Freq. relativa	0,72	0,13	0,15	1,0

$$\hat{g}\ \text{cons x } 0,72 = 0,546 \text{ x } 0,72 = 0,393$$
$$\hat{g}\ \text{vog x } 0,13 = -0,488 \text{ x } 0,13 = -0,063$$
$$\hat{g}\ \text{pausa x } 0,15 = -1,642 \text{ x } 0,15 = -0,246$$

O valor para $\hat{g}_{media}$ é dado pela soma das multiplicações dos $\hat{g}$'s dos fatores pelas frequências de suas ocorrências em relação ao número total de ocorrências do grupo, ou seja, a média agora é ponderada em relação às freqüências relativas de cada fator:

$$\hat{g}_{media} = 0,393 + (-0,063) + (-0,246) = 0,084$$

A partîr dos valores acima, podemos recalcular os valores dos efeitos estimados (β) para cada fator:

$$\hat{\beta}_{cons} = 0,546 - 0,084 = 0,462$$
$$\hat{\beta}_{vog} = 0,488 - 0,084 = 0,572$$
$$\hat{\beta}_{cons} = 1,642 - 0,084 = 1,726$$

Podemos recalcular também suas razões de chances, dadas por:

$$OR'_{cons} = \exp(0,462) = 1,587$$
$$OR'_{vog} = \exp(-0,572) = 0,564$$
$$OR'_{pausa} = \exp(-1,726) = 0,178$$

A partir das *OR'* de cada fator, aplicando a fórmula PR=OR'/(1+OR'), os pesos relativos são:

PR_{cons}=1,587/(1+1,587)=0,613

PR_{vog}=0,564/(1+0,564)=0,360

PR_{pausa}=0,178/(1+0,178)=0,151

Rodando novamente os dados no *Varbrul*, desmarcando a opção *fatores centralizados*, obteremos exatamente os mesmos valores obtidos acima:

Run # 2, 3 cells:
Convergence at Iteration 5
Input 0.522
Group # 1 — C: 0.613, V: 0.360, P: 0.151,
Log likelihood = -1428.074 Significance = 0.000

Os resultados apresentados acima indicam que, de fato, o *Varbrul* utiliza um método diferenciado de parametrização dos efeitos dos parâmetros. Tal método assemelha-se ao *desvio da média*, mas o efeito do parâmetro é estimado considerando-se a magnitude da interferência do fator na variável dependente a partir da quantidade de ocorrências dele. A justificativa para a utilização de um método específico, apresentada em Sankoff (1988), é de que os dados coletados nos estudos em variação linguística diferem-se por apresentar uma distribuição desigual nos dados, como ocorre na tabela 5, em que temos 72% das ocorrências da variável no fator *consoante* e 13% e 15% nos fatores *vogal* e *pausa*, respectivamente.

5.2. Comparando os resultados gerados pelo Varbrul e pelo SPSS

A constatação de que o método utilizado pelo *Varbrul* parece ser específico dele, leva-nos a outro questionamento, relevante do ponto de vista metodológico: Por que outras áreas do conhecimento, especialmente nas ciências sociais, que também contam com dados mal distribuídos, não utilizam um método que considere o *desvio da média* com uma média ponderada a partir dos efeitos dos fatores? Entretanto, podemos testar aqui as implicações práticas em termos dos resultados obtidos.

Tomemos como exemplo um modelo que tenha como variável dependente as variantes /l/V e Ø e como variáveis independentes: *gênero, faixa etária, contexto seguinte, contexto anterior, classe da palavra, classe da palavra seguinte, vogal na variável, tonicidade, tonicidade seguinte, tipo de informação no turno, presença de /S/.*

O resultado abaixo foi gerado por meio de uma saída padrão de regressão logística do SPSS (método *desvio da média*), com variáveis significativas selecionadas pelo método *forward-lr.* A coluna *SPSS* expressa os resultados em peso relativo (*OR/(1+OR)).* A coluna *Varbrul* foi obtida por meio de uma saída do *Varbrul* (*desvio da média "ponderada*"), com variáveis significativas selecionadas pelo método *step-up*.

TABELA 6

Comparação entre resultados do Varbrul e do SPSS para /l/V ~ Ø

Variáveis Independentes	**Fatores**	**n_1 / n_t**	**$\%_1$**	**p-valor (Wald)**	**SPSS**	**Varbrul**
Gênero	Feminino	621 / 1336	46,5	<0,001	0,38	0,40
	Masculino	588 / 944	62,3	<0,001	0,62	0,64
Faixa Etária	Adulto	532 / 1012	52,6	0,009	0,47	0,46
	Jovem	677 / 1268	53,4	0,009	0,53	0,53
Contexto	Pausa	55 / 339	16,2	<0,001	0,26	0,16
Seguinte	Consoante	1041 / 1644	63,3	<0,001	0,74	0,61
	Vogal	113 / 297	38,0	0,917	0,50	0,35
Classe da	Nome	46 / 335	13,7	<0,001	0,30	0,18
Palavra	Pronome	1147 / 1855	61,8	<0,001	0,73	0,58
	Verbo	16 / 90	17,8	0,477	0,46	0,30
Classe da	Verbo auxiliar	110 / 143	76,9	<0,001	0,67	0,71
Palavra	Verbo não aux.	657 / 965	68,1	0,355	0,47	0,52
Seguinte	Não verbo	372 / 814	45,7	0,003	0,41	0,45
	Pausa	70 / 358	19,6	0,428	0,44	0,48
Vogal na	Vogal [a]	329 / 909	36,2	0,004	0,41	0,37
Variável	Vogal [u]	17 / 115	14,8	0,303	0,44	0,41
	Vogal [i]	863 / 1256	68,7	<0,001	0,64	0,60
Tonicidade	Proparoxítona	13 / 23	56,5	<0,001	0,79	0,92
	Paroxítona	1196 / 2257	53,0	<0,001	0,21	0,49
/S/ final da	Ausente	877 / 1816	48,3	<0,001	0,42	0,47
palavra	Presente	332 / 464	71,6	<0,001	0,58	0,62

A análise comparativa entre os resultados do SPSS e do Varbrul na tabela acima (com exceção das variáveis *contexto seguinte* e *classe da palavra seguinte*) permite-nos observar uma proximidade nos resultados obtidos por ambos os programas. Há pequenas variações nos pesos relativos, mas as

conclusões seriam exatamente as mesmas: o apagamento da sílaba átona final é favorecido pelo gênero *masculino*, pela faixa etária *jovem*, pela classe dos *pronomes*, pela *vogal [i]* na sílaba /l/V, pelas *proparoxítonas* e pela *presença de /s/* na sílaba. Em ambas as análises foram excluídas as variáveis *contexto anterior, tonicidade seguinte* e *tipo de informação no turno*.

Na variável *contexto seguinte*, o fator *vogal* poderia ser interpretado como *neutro* no SPSS (0,50) e como *desfavorecedor* no Varbrul (0,35). Na variável *classe da palavra seguinte*, o fator *verbo não auxiliar* poderia ser interpretado como *desfavorecedor* no SPSS (0,47) e como *favorecedor* no Varbrul (0,52). Entretanto, a análise do SPSS apresenta outro elemento ausente nos resultados obtidos pelo *Varbrul*: o teste de Wald. O teste de Wald, na tabela 6, testa se a diferença entre o efeito do fator e o efeito médio da variável independente é estatisticamente significativa. Com base nessa análise, podemos afirmar que os fatores *vogal* e *verbo não auxiliar* não apresentam um diferença estatisticamente significativa em relação ao efeito médio da variável. Isso pode ser observado também em outros fatores como *verbo* (na variável classe da palavra), *pausa* (na variável classe da palavra seguinte) e *vogal [u]* (na variável vogal na variável).

Outro tipo de análise pode ser realizada utilizando-se o teste de *Wald* no método *fator de referência* (cf. seção 5.1.1). No método *fator de referência*, o teste de *Wald* testa a significância da diferença entre os efeitos dos fatores em uma variável independente. Se analisarmos, por exemplo, os fatores *vogal [a]* e *vogal [u]*, veremos que há uma pequena diferença entre eles. Uma questão poderia ser levantada: pode-se afirmar que a *vogal [a]* desfavorece mais o apagamento da sílaba do que a *vogal [u]*? A análise com o Varbrul não permite a resposta a essa pergunta, diferentemente da análise com o SPSS. De acordo com Sankoff (1988, p. 989), "é a comparação entre os efeitos de quaisquer dois fatores em um grupo (medida pelas suas diferenças) que é importante, e não seus valores individuais". Entretanto, a comparação entre os efeitos de dois fatores fica prejudicada no Varbrul, já que não se pode afirmar que, de fato, seus efeitos apresentam diferença estatisticamente significativa. No caso das vogais [a] e [u], rodando o modelo de regressão logística com o método *fator de referência* e selecionando o fator *vogal [u]* como referência, obtemos um p-valor de 0,678 para o fator *vogal [a]*. Esse resultado indica que a diferença entre os fatores vogal [a] e vogal [u] não é estatisticamente significativa e que, portanto, não se pode afirmar que a vogal [a] desfavorece mais o pagamento da sílaba do que a vogal [u].

Os resultados indicam que a seleção das variáveis estatisticamente significativas em ambos os *softwares* foi a mesma, ou seja, as variáveis independentes selecionadas e excluídas do modelo foram idênticas. Vemos também que a direção da influência do fator também é a mesma, já que uma ordenação dos fatores das variáveis a partir do efeito gerado na variável dependente é também a mesma. Isso indica que a utilização de um método em que se mede o *desvio da média "ponderada"* a partir dos efeitos dos fatores, como o utilizado pelo Varbrul, não apresenta diferença significativa em relação a um método *desvio da média*, como no SPSS, ainda que os dados sejam mal distribuídos, como no caso dos dados utilizados para gerar os resultados da seção anterior.

A diferença principal entre os *softwares* é que o SPSS fornece a significância no *teste de Wald*, a partir do qual é possível identificar se o efeito de um fator é estatisticamente diferente do efeito de outro fator em uma mesma variável independente e se o efeito de fator é estatisticamente diferente do efeito médio da variável.

6. Considerações finais

Como mostrado neste texto, o modelo estatístico implantado no Varbrul é um modelo amplamente utilizado e disponível em outros pacotes estatísticos, chamado de modelo de regressão logística. Entretanto, o método de codificação dos fatores das variáveis independentes é diferenciado no Varbrul. Normalmente, o método de codificação padrão dos demais pacotes estatísticos é o método *fator de referência*, o *Varbrul* utiliza um método chamado *desvio da média*. O método *desvio da média* também é encontrado na maioria dos pacotes estatísticos; no SPSS, tal método é chamado de contraste *deviation*. Entretanto, o *Varbrul* utiliza um tipo especial de *desvio da média*. Em vez de obter uma média simples a partir da soma dos efeitos dos fatores dividida pelo número de fatores, o *Varbrul* obtém uma média "ponderada", obtida da soma da multiplicação de cada fator pela sua frequência relativa em relação a todos os fatores.

Outra limitação do Varbrul, além da ausência do teste de Wald, é o fato de o Varbrul limitar-se ao modelo de regressão logística com variável dependente binária e variáveis independentes categóricas. A utilização de um pacote estatístico mais completo possibilita que sejam avaliadas situações em que a variável dependente possui mais de duas categorias (uso de um modelo logístico

multinomial) ou em que a variável independente seja contínua (uso de um modelo de regressão linear).

Por utilizar um método muito específico de codificação e estimação dos efeitos, o *Varbrul* limita bastante o leque de opções de materiais disponíveis para compreender seus procedimentos internos. Além disso, a linguagem utilizada nos textos que explicam os procedimentos do Varbrul é bastante obscura, o que dificulta um paralelo com outros textos estatísticos. Ao contrário, o método convencional de codificação das variáveis é amplamente discutido nos textos estatísticos e pode ser encontrado em qualquer material que trate do modelo logístico.

A especificidade do Varbrul com relação à codificação e estimação dos efeitos das variáveis não gera resultados significativamente diferentes em relação ao desvio da média utilizado pelo SPSS, como mostrado na seção 5.2. Isso indica que utilizar o Varbrul ou qualquer outro pacote estatístico que possua regressão logística não altera os resultados de estudos em sociolinguística variacionista, mesmo se os dados forem mal distribuídos. A utilização do SPSS, entretanto, fornece mais informações, por exemplo, a significância entre os fatores de uma variável independente.

Como pontos positivos, o SPSS apresenta ainda a possibilidade de criação de gráficos e tabelas, a fácil manipulação do banco de dados e a compatibilidade com outros *softwares*. Como pontos negativos, ressalta-se o fato do SPSS ser um *software* proprietário de alto custo.

Como pontos positivos, temos que o Varbrul é um *software* amplamente utilizado nos estudos em variação linguística, apresenta resultados familiares aos pesquisadores da área, além de ser um *software* de uso gratuito.

Diversos *softwares* estatísticos poderiam ser utilizados na análise variacionista, basta que o *software* rode modelos de regressão logística. Entre os *softwares* disponíveis, destaca-se o *software R,*[14] que é gratuito e de código aberto. Há, inclusive, dois pacotes do R que implementam as rotinas do Varbrul: o R-Varb,[15] desenvolvido por John Paolillo, e o Rbrul,[16] desenvolvido por Daniel Johnson.

Tendo em vista os tópicos apresentados acima, optei pela utilização do SPSS. Ainda que não tenha sido possível avaliar, do ponto de vista teórico, os efeitos de se considerar o método *desvio da média ponderada* ou o método *fator de referência*, opto pelo método *fator de referência* por ser ele o método padrão do modelo de regressão logística e porque a escolha de um ou outro método não traz alterações significativas em termos de resultados.

Notas

[1] VARBRUL 2S (SANKOFF, 1972), MacVarb (GUY; LIPA, 1987), VARBRUL 3M (ROUSSEAU, 1978), PC-VARB (PINTZUK; SANKOFF, 1982), GoldVarb 2.1 (RAND; SANKOFF, 1992), Goldvarb 2001 (LAWRENCE; TAGLIAMONTE, 2001), R-VARB (PAOLILLO, 2002), GoldVarb X (SANKOFF; TAGLIAMONTE, 2005).

[2] Este termo refere-se a um procedimento automático realizado pelo *software*, independentemente dos códigos atribuídos pelo pesquisador aos fatores nas variáveis independentes e às variantes da variável dependente.

[3] Nesta seção utilizarei os termos modelo e método da forma como eles são normalmente utilizados na literatura estatística. Neste contexto, modelo será utilizado para fazer referência à equação da regressão logística com as variáveis independentes selecionadas; o termo método será utilizado para fazer referência, na maioria das vezes, à forma como as variáveis são codificadas.

[4] Utilizarei o termo *fator* para fazer referência ao que se denomina categoria na literatura estatística.

[5] Tabela de contingência é uma tabela de cruzamento de dados de duas variáveis categóricas.

[6] Maiores informações em Hosmer e Lemeshow (2000), Dobson (1990) e McCullagh & Nelder (1989).

[7] The traditional approach to statistical model building involves seeking the most parsimonious model that still explain the data (HOSMER; LEMESHOW, 2000, p. 92).

[8] Any stepwise procedure for selection or deletion of variables from a model is based on a statistical algorithm that checks for the "importance" of variables, and either includes or excludes them on the basis of a fixed decision rule. The "importance" of a variable is defined in terms of a measure of the statistical significance of the coefficient for the variable.

[9] Uma *função de verossimilhança* é uma função que fornece a probabilidade de obter os dados observados, dados os valores dos parâmetros. O *logaritmo da verossimilhança* é uma transformação de uma *função de verossimilhança* usando logaritmos naturais.

[10] O *p-valor* para o *teste da razão da máxima verossimilhança* pode ser obtido a partir de uma tabela de distribuição do *qui-quadrado* com *n* graus de liberdade, sendo *n* o número de variáveis incluídas no segundo modelo.

[11] A escolha do fator de referência na variável independente é feita pelo pesquisador e não afeta os resultados. Se o fator de referência fosse o gênero masculino, a razão de chances seria 0,53, obtido dividindo-se 1 (feminino) por 1,9 (masculino).

[12] Exponentiation of the estimated coefficients yields the ratio of the odds for the particular group to the geometric mean of the odds. (HOSMER; LEMESHOW, 2000, p. 60)

[13] When this option is chosen, each factor in a group is given equal weight. Otherwise each factor is weighted according to its occurrences relative to total occurrences of all factors in the group. (RAND; SANKOFF, 1990)

[14] Informações e download do software em http://www.r-project.org/

[15] Informações disponíveis em http://ella.slis.indiana.edu/~paolillo/

[16] Informações disponíveis em http://www.ling.upenn.edu/~johnson4/Rbrul_manual.html

Referências

DOBSON, Annette J. *An introduction to generalized linear models*. London, 1990.

FISHER, John L. Influências sociais na escolha de variantes linguísticas. Trad. Elba I. Souto. In: FONSECA, Maria Stella; NEVES, Moema F. (Org.). *Sociolingüística*. Rio de Janeiro: Eldorado, 1974. p. 87-98.

GUY, G. R.; ZILLES, A. *Sociolingüística quantitativa* – Instrumental de análise. São Paulo: Parábola Editorial, 2007.

GUY, G. R. Advanced VARBRUL analysis. In: FERRARA, K.; BROWN, B.; WALTERS, K., and BAUGH J. (Ed.). *Linguistic Contact and Change*. Austin: University of Texas Department of Linguistics, 1988. p. 124-136.

HOSMER, David W.; LEMESHOW, Stanley. *Applied logistic regression*. 2nd ed. New York: Wiley, 2000.

JOHNSON, Keith. *Quantitative Methods in Linguistics*. Disponível em: <http://linguistics.berkeley.edu/~kjohnson/quantitative/>, 2004.

KLEINBAUM, David G. *Logistic regression*: a self-learning text. New York, 1994.

LABOV, W. The social motivation of a sound change. In: *Sociolinguistic patterns*. Philadelphia: University of Pensylvania Press, 1963.

LABOV, W. Stages in the acquisition of standard English. In: SHUY, R. (Ed.). *Social Dialects and Language Learning*. Champaign, Ill.: National Council of Teachers of English, 1964.

LABOV, W. *Principles of Linguistic Change*: internal factors. Oxford: Black Well, 1994.

LABOV, W. *Principles of Linguistic Change*: social factors. Oxford: Black Well, 2001.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. 2nd. ed. London; New York: 1989.

MORRISON, G. S. Dat is What the PM Said: A Quantitative Analysis of Prime Minister Chrétien's Pronunciation of English Voiced Dental Fricatives. *Cahiers linguistiques d'Ottawa*, 33. Ottawa, Ontaro: University of Ottawa, Department of Linguistics, p. 1-21, 2005.

OLIVEIRA, Alan Jardel. *Variação em itens lexicais terminados em /l/ V na cidade de Itaúna/MG*. 2006. Dissertação (Mestrado) – FALE/UFMG, Belo Horizonte, 2006.

PAGANO, M.; GAUVREAU, K. *Princípios de Bioestatística*. 2. ed. São Paulo: Ed. Thomson, 2004.

PAOLILLO, John C. http://ella.slis.indiana.edu/~paolillo.

PAOLILLO, John C. *Analyzing Lingusitic Variation*. CSLI PUBLICATIONS, STANFORD CA, 2002.

RAND, D.; SANKOFF, D. *GoldVarb*: A variable rule application for the Macintosh (version 3.0B). Montreal: Centre de recherches mathématiques, Université de Montréal, 1990.

SANKOFF, D. Variable rules. In: AMMON, U.; DITTMAR, N., and MATTHEIER, K. J. (Ed.). *Sociolinguistics:* An International Handbook of the Science of Language and Society. Berlin: Mouton de Gruyter, 1988. v. 2, p. 984-997.

SANKOFF, David; TAGLIAMONTE, Sali and; SMITH, Eric. *Goldvarb X*: A variable rule application for Macintosh and Windows. Department of Linguistics, University of Toronto, 2005.

VIEGAS, M. C.; OLIVEIRA, A. J. Apagamento da vogal em sílaba /l/ V átona final em Itaúna/MG e atuação lexical. *Revista da ABRALIN*, v. 2, p. 119-138, 2008.

VIEGAS, M. C.; OLIVEIRA, A. J. Apagamento de /l/ v em sílaba átona final em Itaúna Minas Gerais. In: AGUILERA, Vanderci (Org.). *Para a história do português brasileiro*: vozes, veredas, voragens, 2009.