Capítulo 2
SOCIOLINGUÍSTICA E ESTATÍSTICA

Os estudos primários desta revisão sistemática da literatura, sobre monotongação de ditongos orais no PB, utilizam a metodologia da Sociolinguística Variacionista, e, considerando que nossa análise inclui uma avaliação crítica da aplicação dessa metodologia, é necessário compreender tanto seus pressupostos teóricos quanto sua aplicação, além de conhecer as ferramentas que possibilitam o tratamento quantitativo dos dados. Portanto, este capítulo apresenta (i) os pressupostos teórico-metodológicos da Sociolinguística Variacionista; (ii) algumas noções introdutórias de estatística e os modelos de regressão empregados na Sociolinguística; e (iii) os softwares que podem ser utilizados para executar modelos estatísticos de regressão.

Este capítulo está dividido em cinco seções. A seção 2.1 mostra uma visão geral da Sociolinguística Variacionista como uma ciência interdisciplinar, tratando, ainda, do seu surgimento. Na seção 2.2, discorremos sobre os pressupostos teóricos da Teoria da Variação e Mudança conforme Weinreich, Labov e Herzog (2006[1968]) e, na seção 2.3, sobre as etapas da metodologia da Sociolinguística Variacionista, considerando os principais avanços, sobretudo, na análise quantitativa dos dados (cf. CEDERGREN; SANKOFF, 1974; ROUSSEAU; SANKOFF, 1978; NARO, 2004; SCHILLING-ESTES, 2007; GUY; ZILLES, 2007; TAGLIAMONTE, 2007; 2012; JOHNSON, 2009; OUSHIRO, 2017). A seção 2.4, por sua vez, apresenta noções introdutórias de estatística e os modelos de regressão linear e logística, com destaque para este último, comumente, empregado pela Sociolinguística Variacionista. A seção

2.5 trata dos softwares que podem ser utilizados para executar, entre outras, a modelagem estatística de regressão logística: o Varbrul (e suas versões), o Rbrul e o R (em sua interface RStudio), quando discutimos a utilização e o *output* desses softwares considerando suas limitações e as vantagens de se utilizar cada um deles. Enfim, na seção 2.6, sintetizamos o capítulo.

2.1 O surgimento da Sociolinguística

A Sociolinguística Variacionista tem como objeto de estudo a língua em uso, na vida cotidiana de uma comunidade, considerando os fatores sociais que se correlacionam a ela (LABOV, 1972). Assim, é uma ciência interdisciplinar que estuda a língua em uso se voltando para a relação língua-sociedade. A Sociolinguística está, estreitamente, relacionada a outras três áreas do conhecimento: (i) à sociologia, quando estuda a organização social do comportamento linguístico em termos de uso e atitudes em relação à língua e aos seus falantes; (ii) à antropologia, na medida que estende a descrição e análise linguística de forma a incluir aspectos da cultura da comunidade de fala em que a língua é usada; e (iii) à geografia linguística ou dialetologia, dado seu interesse em considerar diferenças regionais, uma vez que, comumente, as comunidades de fala escolhidas pelo pesquisador são delimitadas geograficamente (COELHO *et al*., 2012).

Apesar de ter havido estudos anteriores que postulavam uma concepção social da língua, tais como os de Meillet (1921) e Bakhtin (1929), é a partir da década de 1960 que a noção de língua como fato social dinâmico, cuja variação pode ser explicada por fatores sociais, ganha força (COELHO *et al*., 2012). Segundo Chambers, Trudgill e Schilling-Estes (2003, p. 5) o marco inicial da Sociolinguística Variacionista foi o ano de 1963, quando William Labov apresentou seu primeiro estudo sociolinguístico

no encontro anual da *Linguistic Society of America* e publicou *The social motivation of a sound change*, sua dissertação de mestrado, na qual, o autor descreve a variação dos ditongos [aj] e [aw] na comunidade norte-americana da ilha de Martha's Vineyard. Ao analisar a influência da variável social *Identidade linguística dos falantes* na aplicação da regra variável, o autor constata que os habitantes que se identificavam com as tradições locais da ilha e não desejavam ir para o continente, centralizavam a vogal /a/ com mais frequência do que os falantes que almejavam uma vida fora de Martha's Vineyard. Em sua tese de doutorado, *The social stratification of English in New York City*, Labov (1966) realizou vários estudos sociolinguísticos na cidade de Nova Iorque, por meio da coleta de dados de fala. Um desses trabalhos foi a análise do fenômeno variável de apagamento do /r/ em posição pós-vocálica. Investigando o comportamento linguístico de atendentes de três grandes lojas de departamento – Sacks, Macy's e S. Klein –, Labov observou um padrão de uso que estava relacionado ao estrato social a que cada loja pertencia. Os informantes das lojas de classe alta e média apresentaram um índice mais alto de manutenção de /r/ do que os da loja de classe baixa. Esses estudos foram, posteriormente, reunidos em sua obra *Sociolinguistic Patterns* publicada em 1972. Ao mostrar a correlação de padrões sociais com a distribuição de uma variável linguística, o autor corrobora sua ideia de que não pode existir uma teoria linguística que não seja social e que o objeto da linguística deve ser a fala, ou seja, a língua como é usada na vida cotidiana de uma comunidade, sem deixar de considerar os fatores sociais que, a ela se correlacionam.

2.2 A Teoria da Variação e Mudança

Na obra *Empirical foundations for a theory of language change* – posteriormente traduzida para o português como *Fundamentos*

empíricos para uma teoria da mudança linguística (WEINREICH; LABOV; HERZOG, (2006[1968]) – Uriel Weinreich, William Labov e Marvin Herzog (1968) desenvolveram a Teoria da Variação e Mudança uma das principais perspectivas da Sociolinguística Variacionista. O objetivo dos autores era descrever a língua como um objeto constituído de heterogeneidade ordenada e apresentar as bases de uma teoria da mudança linguística capaz de fornecer descrições mais adequadas da competência linguística, além de superar os paradoxos que as teorias estruturais, fundamentadas no axioma da homogeneidade, vinham trazendo à linguística histórica (WEINREICH; LABOV; HERZOG, (2006[1968], p. 33).

Weinreich, Labov e Herzog (2006[1968]) desenvolvem algumas críticas aos trabalhos de Hermann Paul (1880), Saussure (1916), Bloomfield (1933) e Chomsky (1965), com base nas quais propõem sua teoria a respeito da mudança linguística. Paul (1880) isola a língua do indivíduo – que Weinreich e seus colegas denominarão idioleto – como o mais legítimo objeto de estudo linguístico, já que, assim, encerraria a natureza estruturada da língua, a coerência do desempenho falado e a regularidade da mudança. Segundo Weinreich, Labov e Herzog (2006[1968], p. 39), ao isolar a língua do indivíduo do uso linguístico do grupo, Paul desenvolveu a dicotomia entre o individual e o social que continua existindo na base das teorias do século XX sobre a mudança linguística.

Os dialetos eram concebidos, por Paul, como agrupamentos de idioletos idênticos e a mudança dialetal consistiria simplesmente em idioletos mudando em paralelo, e a divisão dialetal, não seria mais do que idioletos mudando de forma divergente. Há um pressuposto de homogeneidade tanto para o idioleto quanto para o dialeto. Assim, para tratar da heterogeneidade no interior de um idioleto ou dialeto, Paul utiliza a noção de empréstimo, mas sem explicar os mecanismos desse processo (WEINREICH; LABOV; HERZOG, 2006[1968], p. 54).

Quanto ao trabalho de Saussure (1916), expresso no *Curso de linguística geral*, Weinreich, Labov e Herzog (2006[1968], p. 55) retomam a antinomia entre sincronia e diacronia afirmando que, para Saussure a sistematicidade da língua depende da existência, dentro do indivíduo, de uma faculdade de associação e coordenação. As relações entre elementos de uma língua se localizam na consciência do falante e, para garantir a realidade psicológica do seu objeto de estudo, Saussure estabelece que tal objeto seja homogêneo, dessa forma, o teórico recorta um objeto sincrônico, sempre homogêneo, e, de acordo com Weinreich, Labov e Herzog, não há nada em sua teoria que possa acomodar a língua como heterogeneidade ordenada. Faltam recursos teóricos para tratar da coexistência entre formas conservadoras e inovadoras num mesmo falante, bem como a coexistência de sistemas de dialetos vizinhos na competência dos mesmos falantes. Tendo isso em vista, os autores consideram que apesar de Saussure ter desempenhado um papel revolucionário na história da linguística, não foi além de Paul em sua capacidade de lidar com a língua como fato social, já que, para ele a condição para lidar com a língua como fenômeno social era sua completa homogeneidade. Desse modo, sua teoria não contribuiu, de nenhuma forma, com o estudo da língua como heterogeneidade ordenada. Weinreich, Labov e Herzog (2006[1968], p. 58) chegam à mesma conclusão quanto à linguística norte-americana (Bloomfield, 1933), já que, apesar de haver um interesse pela diversidade dentro de uma comunidade de fala não se chega a uma descrição sistêmica dessa diversidade.

Quanto à linguística chomskyana, os autores apontam que o único objeto legítimo de análise, para essa teoria linguística, seria um sistema homogêneo, uma vez que a mesma se ocupa de um falante-ouvinte ideal, numa comunidade de fala completamente homogênea. Como nos trabalhos de Paul, Saussure e Bloomfield, não há, na proposta chomskyana, procedimentos para ultrapassar

a diversidade observada no comportamento linguístico, além disso, Chomsky, em consonância com a posição dos fundadores da linguística geral moderna, declara que tal diversidade é teoricamente irrelevante e, na sequência, afirma que não se tinha oferecido nenhuma razão convincente para modificar essa posição. Weinreich, Labov e Herzog (2006[1968], p. 60) discordam dessa declaração e afirmam que os desvios de um sistema homogêneo não são, todos, erros aleatórios de desempenho, como apontado pela teoria de Chomsky, tais desvios são codificados num alto grau, ou seja, a heterogeneidade da língua é codificada em alto grau e está integrada na competência linguística do falante.

Em resumo, para Paul, Saussure, Bloomfield e Chomsky variabilidade e sistematicidade se excluíam mutuamente, o que Weinreich, Labov e Herzog (2006[1968]) apontam como sendo um paradoxo das teorias baseadas no axioma da homogeneidade. Em tais teorias, a estrutura era associada à homogeneidade e à funcionalidade da língua, entretanto, se uma língua deveria ser estruturada para funcionar de forma eficiente, como as pessoas continuam a falar enquanto a língua muda, isto é, enquanto passa por períodos de menor sistematicidade? Por que não se observa ineficiências na prática? De acordo com Weinreich, Labov e Herzog essa é a questão fundamental com a qual a teoria da mudança linguística precisa lidar, e a solução está no rompimento da identificação da estruturalidade com a homogeneidade. Os autores defendem que uma explicação razoável da mudança depende da possibilidade de descrever a diferenciação ordenada dentro da língua, posto que toda mudança implica variabilidade e heterogeneidade. Além disso, argumentam que o domínio de um falante nativo de estruturas heterogêneas é parte da sua competência linguística monolíngue e que numa língua, que serve a uma comunidade complexa, a ausência de heterogeneidade estruturada seria disfuncional (WEINREICH; LABOV; HERZOG, 2006[1968], p. 36).

Dentro de uma comunidade de fala há formas distintas da mesma língua que coexistem. Essas formas podem ser denominadas "estilos", "padrões", "gírias", "jargões", "jeito antigo de falar", ou "variedades funcionais" e compartilham as seguintes propriedades: (i) oferecem meios alternativos de dizer a mesma coisa, isto é, fornecem a mesma informação referencial; e (ii) estão conjuntamente disponíveis a todos os membros da comunidade de fala. Alguns falantes podem não ser capazes de produzir enunciados em todas as variedades disponíveis com igual competência devido a restrições em seu conhecimento pessoal, além de práticas associadas ao seu status social, mas todos os falantes, geralmente, tem capacidade de interpretar enunciados em qualquer uma dessas variedades, bem como compreender a escolha de qualquer variedade por outro falante (WEINREICH; LABOV; HERZOG, 2006[1968], p. 96).

Weinreich, Labov e Herzog (2006[1968], p. 99) apontam que, ao longo das décadas de 1920 e 1930, houve uma tendência, por parte dos linguistas, a se afastar da unidade do idioleto, postulado por Paul, como objeto de estudo, afirmando que tais estudos confirmam "o modelo de um sistema ordenadamente heterogêneo em que a escolha entre alternativas linguísticas acarreta funções sociais e estilísticas, um sistema que muda acompanhando as mudanças na estrutura social". Um desses trabalhos é o de Mathesuis e seus colegas em Praga que usaram uma abordagem multiestratificada para caracterizar sistemas coexistentes na mesma comunidade (WEINREICH; LABOV; HERZOG, 2006[1968], p. 96).

Conforme Weinreich, Labov e Herzog (2006[1968]), tal concepção multiestratificada da língua pode ser utilizada com fins puramente analíticos, para representar a língua como um "diassistema" composto por dialetos-membros, mas, para que essa teoria tenha importância, também, na linguística histórica, isto é, para o pesquisador que investiga a mudança linguística, os autores ressaltam que os estratos que a língua inclui, ainda que

funcionalmente distintos, devem estar funcionalmente disponíveis a um grupo de falantes, visto que, é somente quando dois ou mais dialetos estão conjuntamente disponíveis a um grupo, que alterna entre eles, que a formulação multiestratificada é relevante para se entender a mudança linguística. Ademais, os autores afirmam insistir na distintividade funcional por duas razões: (i) os estratos devem estar em competição e não em complementariedade; (ii) é preciso fornecer uma descrição rigorosa das condições que governam a alternância dos sistemas. As regras devem incluir fatores extralinguísticos como ambientes condicionadores além de "fornecer uma descrição linguística das relações que governam unidades igualmente presentes ao longo dos extratos" (WEINREICH; LABOV; HERZOG, 2006[1968], p. 100). Os resultados do estudo de Lambert e seus colegas (1960, 1967) são retomados para mostrar que a escolha ou alternância entre subsistemas pode ser determinada por atitudes sociais, mais especificamente, por traços de personalidade, inconscientemente, atribuídos a falantes dessa variedade, e tais atitudes podem levar ao seu desenvolvimento ou obsolescência (WEINREICH; LABOV; HERZOG, 2006[1968], p. 102).

Em suma, o caráter heterogêneo dos sistemas linguísticos seria o produto de combinações, alternâncias ou mosaicos de subsistemas distintos, conjuntamente disponíveis, e cada subsistema "é concebido como um corpo coerente e integral de regras do tipo categórico, neogramático: o único aparato teórico adicional necessário é um conjunto de regras que afirmem as condições para a alternância" (WEINREICH; LABOV; HERZOG, 2006[1968], p. 102). Weinreich, Labov e Herzog afirmam que não seria possível abstrair um dos subsistemas desse complexo sem perder informações necessárias à análise linguística, e acrescentam que os diversos estudos que isolam um desses vários sistemas teriam sido desenvolvidos sob a suposição de que um sistema homogêneo, invariante, seria a única base possível para a descrição linguística, não oferecendo, assim,

nenhuma base racional para a explicação da mudança linguística, na medida em que, para que um estudo seja capaz de tal coisa, deve ser baseado num modelo de língua diferenciada, e incluir elementos variáveis dentro do próprio sistema. Para esses autores a associação entre estrutura e homogeneidade é uma ilusão já que a "estrutura linguística inclui a diferenciação ordenada dos falantes e dos estilos através de regras que governam a variação na comunidade de fala" (WEINREICH; LABOV; HERZOG, 2006[1968], p. 125).

Com a finalidade de explicar a complexidade da estrutura linguística nesse modelo diferenciado, de heterogeneidade ordenada, Weinreich, Labov e Herzog introduzem o conceito de variável linguística, um elemento linguístico que varia, dentro do sistema, sendo constituído por duas ou mais variantes (suas possíveis realizações) e controlado por uma regra. Conforme os autores, a variável linguística deve ser estabelecida sob condições estritas, para que seja parte da estrutura linguística. Uma condição necessária para admitir a unidade estrutural seria a "evidência quantitativa para a covariação entre a variável em questão e algum outro elemento linguístico ou extralinguístico" (WEINREICH; LABOV; HERZOG, 2006[1968], p. 107). Os autores propõem a seguinte representação para as regras:

(1) /A/ → g[B] / X __ Y
(2) g[B] = f (C, D, E...)

Em que B é um ou mais traços de A, a expressão g[B] é a variante linguística definida pela regra e C, D, e E são variáveis linguísticas ou extralinguísticas. Em (1) tem-se que o traço B é elidido nos contextos em que aparece entre X e Y. Em 2 tem-se que g[B] é uma função de C, D e E.

Para Weinreich, Labov e Herzog (2006[1968], p. 124-125), uma mudança começa quando um dos traços, que estavam em variação

na fala, se difunde através de um subgrupo da comunidade de fala e assume uma significação social. Quando a mudança é encaixada na estrutura linguística, ela é gradualmente generalizada a outros elementos do sistema, por conseguinte, a mudança não á algo que ocorre de forma instantânea e sim gradual. A concretização da mudança linguística e a seleção de uma das alternativas como uma constante são acompanhadas pela perda de significação social que o traço, que está desaparecendo, possuía. Desse modo, segundo Weinreich, Labov e Herzog (2006[1968], p. 121-125), a Teoria da Variação e Mudança se propõe resolver cinco problemas empíricos:

1. Determinar as condições para a mudança, ou seja, os fatores condicionantes, haja vista que algumas mudanças só ocorrem sob determinadas condições.
2. Descrever o processo de transição da mudança que se dá (i) “à medida que um falante aprende uma forma alternativa”, (ii) “durante o tempo em que as duas formas existem em contato dentro de sua competência”, e (iii) “quando uma das formas se torna obsoleta” (WEINREICH; LABOV; HERZOG, 2006[1968], p. 122).
3. Explicar o encaixamento dos traços mutantes na estrutura linguística e na comunidade de fala, visto que, a mudança raramente é um movimento de um sistema inteiro para outro. Em vez disso, um conjunto limitado de variáveis [5], num sistema, altera seus valores modais, gradualmente, de um polo para outro e os traços mutantes são encaixados na estrutura linguística. Como a estrutura linguística mutante está encaixada na comunidade de fala, as variações sociais e geográficas são elementos intrínsecos a estrutura e, portanto,

5 O conceito da variável como um elemento estrutural torna desnecessário tratar variações no uso como algo externo ao sistema, já que, o controle dessas variações faz parte da competência linguística dos falantes (WEINREICH, LABOV; HERZOG,2006[1968], p. 123).

fatores sociais influenciam o sistema como um todo, no entanto, a significação não é equitativamente distribuída por todos os fatores, isto é, cada fator possui uma significação distinta (WEINREICH; LABOV; HERZOG, 2006[1968], p. 123).

4. Estabelecer, empiricamente, os correlatos subjetivos, das avaliações que os falantes fazem dos diversos estratos e variáveis contidas numa estrutura heterogênea.
5. Explicar a implementação da mudança linguística, o que implica descrever as motivações linguísticas e sociais para a mudança.

Isto posto, na próxima seção discorremos sobre a metodologia de análise da Sociolinguística Variacionista.

2.3 A Sociolinguística Variacionista

Esta seção está dividida em duas subseções, a subseção 2.3.1 discute as etapas da metodologia da Sociolinguística Variacionista, desde a identificação da variável linguística a ser analisada, passando pelos critérios de seleção dos informantes, pela escolha da comunidade de fala e pelo trabalho de campo, até o tratamento quantitativo dos dados conforme Labov, (1972), Silva, 2004b, Schilling-Estes, 2007; Guy e Zilles, 2007, Tagliamonte (2007; 2012) e Freitag (2016) e a interpretação dos resultados. Enfim, a subseção 2.3.2 apresenta a evolução do modelo estatístico empregado na análise quantitativa dos dados, desde a década de 1960 até os dias atuais (cf. CEDERGREN; SANKOFF, 1974; ROUSSEAU; SANKOFF, 1978; NARO, 2004; TAGLIAMONTE, 2012; JOHNSON, 2009; OUSHIRO, 2017).

2.3.1 A metodologia da Sociolinguística Variacionista

Um modelo quantitativo é entendido como um construto linguístico que procura explicar a realização de diferentes variantes linguísticas, estendendo-se de modo a explicar, também, os padrões quantitativos de uso dessas variantes por meio de um modelo matemático (GUY; ZILLES, 2007, p. 101). Mais adiante discorremos sobre o modelo matemático utilizado nas análises sociolinguísticas, como as realizadas nas dissertações que compõem nosso *corpus*, mas, por ora, focaremos nas etapas da metodologia da Sociolinguística Variacionista. De acordo com Labov (1972) essa metodologia é composta por várias etapas, mas é possível resumi-las em cinco:

1. Identificação da variável linguística (resposta) e das possíveis variáveis previsoras (fatores que possam influenciar a escolha de uma das variantes da variável resposta);
2. Seleção da comunidade de fala e dos informantes;
3. Coleta de dados (trabalho de campo);
4. Análise quantitativa e apresentação dos dados;
5. Interpretação dos resultados e análise dos fatores que influenciam o uso de uma das variantes da variável resposta.

A identificação de uma variável linguística (resposta) consiste na definição do objeto de estudo e implica definir suas variantes, isto é, suas possíveis realizações – ou formas diferentes de dizer uma mesma coisa. Quando a variável possui apenas duas variantes, geralmente, se propõe uma forma subjacente e uma regra variável que a converte, no curso de uma derivação, em uma forma superficial diferente. Quando essa regra é aplicada ocorre a variante que difere da forma subjacente e quando não é aplicada realiza-se a forma que corresponde à estrutura subjacente (GUY;

ZILLES, 2007, p. 36-38). Após a definição da variável linguística e suas variantes, bem como a formulação de uma regra variável, é preciso identificar os fatores que possam influenciar a aplicação da regra. Neste ponto, o conhecimento de como a língua funciona conduzem o pesquisador na elaboração de hipóteses razoáveis de investigação (GUY; ZILLES, 2007, p. 36-38).

A segunda etapa consiste na seleção da comunidade de fala, na qual a pesquisa será desenvolvida, mas antes de pensar em tal escolha, é importante fazer algumas reflexões sobre o que se entende por comunidade de fala. Por décadas, sociolinguistas se pautaram na definição clássica de Labov (1966) de comunidade de fala como um grupo de pessoas que compartilha normas e formas de avaliação de variáveis linguísticas comuns, o que não significa que todos os membros de uma comunidade falem de forma idêntica, mas que todos seriam guiados pelas mesmas normas de fala, isto é, todos teriam o mesmo entendimento sobre qual seria a forma de falar. Entretanto, alguns pesquisadores questionam esse fato indicando a possibilidade de haver um espaço para conflitos sociais e linguísticos dentro de uma comunidade (SCHILLING-ESTES, 2007, p. 167), ou seja, seria possível que nem todos os seus membros compartilhem, exatamente, as mesmas normas.

Há, ainda, questionamentos quanto à dimensão de uma comunidade de fala que, comumente, pode ser uma área geograficamente delimitada (cf. LABOV, 1966), como um bairro, uma cidade, um estado etc. Mas um país, por exemplo, poderia ser uma comunidade de fala? Schilling-Estes (2007, p. 167) problematiza, justamente, que "área geográfica" configure um critério para a definição primária de comunidade de fala. Para tanto, a autora apresenta alguns questionamentos como, por exemplo, um grupo que conversa regularmente, através da internet, sobre uma área de interesse comum, não poderia ser considerado uma comunidade de fala? A mesma autora aponta que alguns pesquisadores, interessados

em grupos menores e seus padrões interacionais, frequentemente, esclarecem que suas análises são baseadas no estudo de "redes sociais" ao invés de uma comunidade de fala geograficamente delimitada, enquanto outros pesquisadores, cujo foco é estudar não apenas padrões interacionais, mas também, as pessoas e suas práticas sociais – como, por exemplo, entender como essas práticas moldam e são moldadas por seus usos linguísticos – preferem trabalhar com uma "comunidade de prática". A despeito da complexidade das questões apontadas, Schilling-Estes (2007, p. 166-167) afirma que não há um melhor tipo de comunidade de fala para se estudar – muitos pesquisadores têm se beneficiado do estudo de todos os tipos de comunidades – e que a escolha do tipo de comunidade depende do interesse do pesquisador. Os estudos sociolinguísticos incluídos nesta revisão sistemática foram baseados na análise de dados de comunidades de fala geograficamente delimitadas, em sua maioria, municípios brasileiros.

Considerando que a população das comunidades escolhidas, geralmente, é muito numerosa, o pesquisador precisa selecionar os informantes que constituirão a amostra (cf. SCHILLING-ESTES, 2007). Conforme Silva (2004b, p. 119-120), para que os resultados do estudo sejam representativos de toda a população, é preciso levar em consideração que o número de informantes selecionados vai depender: (i) da homogeneidade da população quanto à faixa etária, classe social, escolaridade etc.; (ii) do número de variáveis consideradas no estudo; (iii) do fenômeno estudado, haja vista que, a língua é mais homogênea para alguns fenômenos do que para outros; e (iv) do método de análise. Assim, a amostra pode ser reduzida de acordo com a precisão do método estatístico empregado.

Definido o número de informantes que constituirão a amostra é preciso selecioná-los. Contudo, é fundamental garantir que a amostra seja estatisticamente representativa, permitindo a realização

de inferências estatísticas, isto é, que os padrões observados na amostra possam ser generalizados para a população como um todo. Um princípio básico para garantir a representatividade dos dados é utilizar uma amostra aleatória de modo a dar a cada informante da população a mesma probabilidade de serem incluídos na amostra (cf. SILVA, 2004b; GUY; ZILLES, 2007; SCHILLING-ESTES, 2007). Essa seleção aleatória pode ser realizada por diferentes métodos. Aqui apresentaremos dois: (i) método aleatório simples e (ii) método aleatório estratificado.

No primeiro método, os informantes devem ser sorteados, de forma que todos tenham a mesma chance de serem selecionados. Para utilizar esse método é necessário que a amostra seja muito grande, a fim de incluir todos os estratos da comunidade, e que a população seja muito homogênea. O desafio imposto por este método é a necessidade de se ter acesso aos dados de todos os membros da comunidade. Ademais, na maioria dos casos, o pesquisador está interessado na relação entre a variação linguística e as características sociais específicas como gênero, faixa etária, classe social etc. e não há garantia de que a amostra incluirá membros de todos os grupos sociais de interesse ou que cada grupo será representado equitativamente. Logo, se o pesquisador pretende analisar um comportamento linguístico de acordo com a classe social, por exemplo, e a comunidade de fala é formada, majoritariamente, por cidadãos de classe baixa, é provável que a amostra não inclua um número suficiente de informantes de cada classe (cf. SILVA, 2004b; SCHILLING-ESTES, 2007).

No método aleatório estratificado a população é dividida em estratos sociais – também chamados células ou casas – cada uma composta por informantes com as mesmas características sociais, sendo que a seleção dos informantes para preencher cada célula deve ser aleatória. Tal método possibilita que o pesquisador obtenha um número suficiente de informantes de todos os estratos que

pretenda analisar (Silva, 2004b). Dessa forma, se a única variável social a ser analisada é *Gênero*, a amostra poderia ser formada por 10 informantes, com 5 mulheres numa casa e 5 homens na outra. Mas, se acrescentarmos uma variável como *Nível de Escolaridade* com três níveis (ensino fundamental, ensino médio e ensino superior) precisaríamos das seguintes células:

Quadro 1: Células de informantes – Exemplo de organização de uma amostra

5 mulheres com ensino fundamental	5 mulheres com ensino médio	5 mulheres com ensino superior
5 homens com ensino fundamental	5 homens com ensino médio	5 homens com ensino superior

Desse modo, para saber o tamanho da amostra, basta multiplicar o número de células pelo número ideal de informantes em cada célula. No exemplo do quadro 1, a amostra deveria ser constituída por 30 (6x5) informantes. Porém, para que o pesquisador seja capaz de determinar as categorias sociais importantes para a população é importante entender a relação entre significados linguísticos e sociais na comunidade que está sendo estudada. Segundo Schilling-Estes (2007, p. 170-171),

> More and more, variationists are realizing that the best studies are not fully planned in advance but rather that one achieves the fullest understanding of the interrelation between linguistic and social meanings if one keeps an open mind and allows the particularities of each different community, as well as community members' own perspectives, to inform studies as they progress. In other words, variationists increasingly are seeking to use ethnographic methods involving careful, long-term participation in and observation of the communities they study rather than relying solely on pre-determined, "objective" social factors, whether the population under study is a small community of practice that may not be immediately evident to outsiders.

Com os informantes selecionados, a partir das categorias sociais estabelecidas, a próxima etapa é a coleta de dados, comumente, realizada por meio de gravações de entrevistas sociolinguísticas. A entrevista é o procedimento mais utilizado em estudos sociolinguísticos e consiste na interação entre um informante e o próprio pesquisador – ou alguém que trabalhe com o pesquisador – devendo ser o mais informal possível.

Conforme Labov (2008[1972], p. 63) a entrevista é um método básico para obter uma grande quantidade de dados de fala confiáveis de uma pessoa, contudo, é fundamental que o entrevistado se sinta à vontade de maneira que não se preocupe com seu modo de falar e, assim, seja possível capturar sua fala vernacular. Para Labov (2008[1972], p. 239) "o modo de operação ideal é o linguista se engajar numa conversa normal com o informante e ser capaz de elicitar o uso natural de dada forma sem usá-la ele mesmo".

Destarte, uma entrevista sociolinguística, conforme conceituada, por Labov (2008[1972]; 1984) e Wolfram e Fasold (1974) deve se aproximar, o máximo possível de uma conversação casual. As questões devem ser baseadas em tópicos que sejam de interesse geral na comunidade em estudo e realizadas de forma natural minimizando a atenção do falante para o fato de que está sendo gravado para um estudo linguístico, com a finalidade de obter uma grande quantidade de amostras de fala de um informante que se aproxime o máximo possível de sua fala cotidiana ou vernacular (SCHILLING-ESTES, 2007, p. 171-172). Haja vista que segundo Labov (2008[1972]), o estilo de fala mais regular em sua estrutura e em sua relação com a evolução da língua é o vernacular, no qual o falante dispensa a mínima atenção a sua fala.

A estrutura básica da entrevista sociolinguística tem sido modificada desde a sua idealização. Alguns pesquisadores, em seus estudos, entrevistaram grupos de amigos, ao invés de um participante por vez (cf. LABOV; COHEN; ROBBINS; LEWIS, 1968;

LUCAS; BAYLEY; VALLI; ROSE; WULF, 2001), ou utilizar duplas de entrevistadores, inclusive pares naturais como cônjuges (cf. WOLFRAM; HAZEN; SCHILLING-ESTES, 1999) a fim de evitar a formalidade da entrevista entre um informante e um entrevistado. Outros pesquisadores têm utilizado outras técnicas, realizando, em lugar de uma entrevista, previamente planejada, uma conversação espontânea com o informante (cf. MILROY; MILROY, 1978; CHILDS; MALLINSON, 2004; VIEIRA; BALDUINO, 2020, 2021), o que exige do pesquisador um maior conhecimento da comunidade. Além disso, alguns estudos são baseados em conversações espontâneas ou interações livres entre dois ou mais informantes, sem a presença de um entrevistador (cf. STUART-SMITH,1999; MACAULAY, 2002), o que fornece ao pesquisador dados de fala bastante espontânea, mas, também, uma grande quantidade de sobreposição de falas o que dificulta a transcrição. Esse tipo de gravação é interessante, principalmente, quando se quer analisar a conversação (cf. SILVA, 2004b; SCHILLING-ESTES, 2007).

Após planejar a entrevista ou a técnica a ser utilizada na coleta de dados, o pesquisador precisará realizar o trabalho de campo, propriamente dito, que pode ser numa comunidade familiar, na cidade ou estado em que o pesquisador reside, por exemplo, ou estrangeira, em outro estado ou em outro país. Quando a comunidade é estrangeira, os desafios impostos ao pesquisador são maiores, dado que o pesquisador não conhece a comunidade, nem seus membros, sua cultura e seus costumes, sendo possível, ainda, que ele precise lidar com questões relacionadas a preconceito – racial, de gênero, entre outros (cf. SCHILLING-ESTES, 2007).

Antes de iniciar o trabalho de campo, o pesquisador precisará (i) fazer o primeiro contato, na comunidade, com uma pessoa que o auxilie a encontrar membros da comunidade que se encaixem nos extratos predeterminados, e concordem em participar da pesquisa; (ii) encontrar um local silencioso e adequado para realizar as

gravações; e (iii) seguir os procedimentos éticos básicos necessários quando se faz uma pesquisa com seres humanos, como preservar a confidencialidade da identidade e informações pessoais dos participantes e disponibilizar a estes um Termo de Consentimento Livre e Esclarecido (TCLE) com informações gerais e específicas sobre o estudo, o qual cada participante deve assinar expressando seu consentimento em participar da pesquisa (cf. SCHILLING-ESTES, 2007; FREITAG, 2016).

No Brasil toda pesquisa com seres humanos deve seguir as mesmas regras, não importando sua natureza. Conforme a Resolução 196/96, II.1, e na III.2, do Conselho Nacional de Saúde, "Todo procedimento de qualquer natureza envolvendo o ser humano, cuja aceitação não esteja ainda consagrada na literatura científica, será considerado como pesquisa e, portanto, deverá obedecer às diretrizes da presente resolução". Esses procedimentos incluem, entre outros, aqueles de natureza instrumental, ambiental, nutricional, educacional, sociológica, econômica, física, psíquica ou biológica (farmacológicos, clínicos ou cirúrgicos e de finalidade preventiva, diagnóstica ou terapêutica).

Realizada a coleta de dados, chegamos à quarta etapa, isto é, à análise quantitativa das ocorrências que possibilita o estudo da variação linguística, permitindo ao pesquisador entender "sua sistematicidade, seu encaixamento linguístico e social e sua eventual relação com a mudança linguística" (GUY; ZILLES, 2007, p. 73). Essa parte consiste numa análise distribucional dos dados, através de um método quantitativo, e o cálculo do efeito de variáveis (linguísticas e sociais), na seleção de uma das variantes da variável linguística em estudo (cf. TAGLIAMONTE, 2007).

O objetivo da análise distribucional dos dados é conseguir resumi-los de forma que os detalhes sem importância sejam minimizados e que se apresente uma visão geral do que é relevante para o pesquisador, isso tudo sem distorcer, significativamente,

os dados originais. A escolha do método para resumir os dados depende do tipo da variável a ser estudada: se numérica como valores de formantes para articulação de vogais, ou nominal, como a ocorrência ou não de um segmento fonológico (cf. GUY; ZILLES, 2007; OUSHIRO, 2017).

Na análise de uma variável nominal, como a realização variável de um ditongo oral do PB, o primeiro passo é verificar a frequência das variantes nos dados e em cada um dos contextos considerados, nomeadamente, os níveis, ou fatores, que compõem cada uma das variáveis previsoras, ou independentes (cf. TAGLIAMONTE, 2012). O número de ocorrências que compõem o *corpus* de um estudo sociolinguístico pode variar bastante, o número mínimo de ocorrências, por contexto, geralmente aceito é 30 (cf. CEDERGREN; SANKOFF, 1974; TAGLIAMONTE, 2012). Segundo Tagliamonte (2012, p. 136), "general statistical laws dictate that with fewer than 10 tokens there is a high likelihood [6] of random fluctuation, but with numbers greater than 10 there is 90% conformity with the predicted norm, rising to 100% with 35 tokens".

Na sequência é preciso calcular a proporção, ou percentual, correspondente aos valores de frequência encontrados. A porcentagem varia num intervalo de 0% a 100% e fornece um modo de resumir a proporção de resultados alternativos. A análise distribucional e organização dos dados, em valores de frequência e proporção, possibilita que o pesquisador visualize a distribuição das variantes, verificando e demonstrando as tendências encontradas através da apresentação desses dados por meio de tabelas ou gráficos de modo a facilitar a compreensão do fenômeno estudado (cf. GUY; ZILLES, 2007: TAGLIAMONTE, 2007; TAGLIAMONTE, 2012). Essa análise distribucional possibilita a formulação de hipóteses as quais

6 O termo *likelihood* pode ser entendido aqui como probabilidade.

serão testadas por meio de um modelo estatístico.

Antes de tratar do cálculo do efeito dos fatores (linguísticos e sociais), na escolha de uma variante, é necessário esclarecer o conceito de "regra variável", uma regra de reescrita sensível ao contexto que relaciona um par de variantes como x → y, de forma que, quando a regra se aplica, ocorre "y" e quando não é aplicada ocorre "x" (cf. LABOV, 1969; CEDERGREN; SANKOFF, 1974; GUY; ZILLES, 2007). Segundo Guy e Zilles (2007, p. 49-50) a análise da regra variável envolve "a contagem das ocorrências da variável, a descrição de tendências e da extensão da variabilidade, bem como das restrições ou fatores que a influenciam, mediante métodos estatísticos". Desse modo, a análise de uma regra variável é um tipo de análise multivariada[7], desenvolvida na linguística como uma forma de dar conta da variação estruturada, governada por regras, no uso da língua, cujo objetivo é separar, quantificar e testar a significância dos efeitos de fatores contextuais, sociais e linguísticos, na escolha de uma das variantes da variável linguística em análise (GUY; ZILLES, 2007).

Explicitemos os termos utilizados na análise de uma regra variável: (i) a variável que se está estudando é a variável resposta, chamada, na Sociolinguística Variacionista, de *variável dependente*. Se tal variável for categórica pode ser classificada de acordo com o número de variantes que possui, podendo ser binária, quando possui duas variantes, ternária, com três variantes ou eneária, com mais de três; (ii) as variáveis linguísticas (como *Contexto fonológico precedente* e *Tonicidade da sílaba*) e sociais (como *Sexo* ou *Faixa etária* dos informantes) que influenciam a variável dependente favorecendo ou desfavorecendo, em algum grau, a aplicação de

7 Uma análise multivariada é um tipo de análise que – diferentemente de uma análise univariada que verifica o efeito de apenas uma variável independente – testa o efeito de mais de uma variável independente sobre uma variável dependente, incorporando a ideia de que processos linguísticos são influenciados, simultaneamente, por diversas variáveis independentes, linguísticas e sociais (cf.: GUY; ZILLES, 2007).

uma regra variável, são as variáveis previsoras, denominadas na Sociolinguística, *variáveis independentes* ou *conjunto de fatores*. É importante destacar que cada regra variável só é capaz de modelar com sucesso uma única variável resposta, que possua apenas duas variantes possíveis, dessa forma, se houver mais de duas variantes, é preciso postular regras adicionais (cf. GUY; ZILLES, 2007).

Cada variável independente é constituída por um conjunto de níveis (fatores) como as categorias *verbo* e *nome* da variável *Classe gramatical da palavra*. Segundo Guy e Zilles (2007, p. 38), cada grupo de fatores pode ser definido "como um locus na regra variável onde ocorre o condicionamento e consiste em uma lista exaustiva de todos os possíveis fatores mutuamente exclusivos que podem ocorrer naquele locus". Assim, cada fator é um possível valor de uma variável independente.

A fim de exemplificar o que foi exposto acima, pensemos numa análise da variação de um ditongo oral como [ow], por exemplo. A variável resposta (dependente) seria o par de variantes – ou possíveis realizações – do ditongo, nomeadamente, o ditongo propriamente dito [ow], e a vogal [o]. A regra que controla a variável dependente é a monotongação que, quando é aplicada, reduz o ditongo a uma vogal simples, caso contrário o ditongo é realizado integralmente. As variáveis independentes (previsoras), por sua vez, são organizadas em duas categorias: (i) linguísticas – como *Classe gramatical da palavra* e *Tonicidade da sílaba* em que o ditongo está contido; e (ii) sociais – como *Faixa etária* e *Nível de escolaridade* dos informantes. Cada uma dessas variáveis é constituída por níveis (fatores, desse modo, a variável linguística Tonicidade da sílaba pode ser composta pelos fatores: *átona* e *tônica*, já os fatores da variável social *Nível de escolaridade* dos informantes, podem ser, por exemplo, *ensino fundamental*, *ensino médio* e *ensino superior*, sendo que, cada um desses fatores possui um efeito sobre a aplicação da regra de monotongação.

Para que a análise de uma regra variável seja realizada com sucesso as variáveis independentes ou grupos de fatores devem ser ortogonais e independentes, isto é, cada fator de um grupo deve ser capaz de coocorrer com cada um dos fatores em todos os outros grupos, além de representar uma restrição logicamente separada e isolável (GUY; ZILLES, 2007, p. 38; TAGLIAMONTE, 2012).

O centro da análise de uma regra variável é a estimativa dos valores dos efeitos dos fatores, sobre a aplicação da regra, o que requer o cálculo de um valor correspondente ao efeito de cada fator de uma variável independente sobre a escolha de uma das variantes da variável linguística analisada. Os valores, do efeito de cada um dos fatores, são estimados por um modelo estatístico, nomeadamente, o modelo logístico, proposto por Henrietta Cedregen e David Sankoff em 1974 e aprimorado por Pascale Rousseau e David Sankoff em 1978, sobre o qual discorreremos na seção 2.4.

Salientamos que, quando se utiliza um método estatístico de análise, a apresentação dos dados é algo fundamental para interpretação dos resultados, última etapa dessa metodologia. Portanto, os dados devem ser sintetizados em tabelas que facilitem a compreensão do fenômeno estudado e permita a interpretação dos efeitos dos fatores que favorecem ou desfavorecem a aplicação da regra variável, bem como a realização de análises futuras.

Em suma, a Sociolinguística Variacionista utiliza métodos estatísticos para verificar como, e em que medida, fatores de variáveis linguísticas e sociais influenciam a aplicação de uma regra variável, que resulta na escolha de uma variante da variável linguística em estudo.

2.3.2 Evolução do modelo matemático utilizado pela Sociolinguística

O objetivo da análise quantitativa empregada pela Sociolinguística Variacionista é verificar o quanto cada variável independente, estrutural ou social, contribui para a realização de uma ou outra forma variante que constitui a variável dependente. Considerando que, na língua em uso, essas variáveis independentes sempre aparecem associadas, a atuação de uma regra variável ocorre em consonância com o efeito simultâneo de mais de uma variável independente. Logo, é necessário calcular o efeito simultâneo de todas as variáveis independentes, presentes em determinado contexto. Assim, o objetivo de uma análise quantitativa da variação é entender o comportamento de uma variável dependente de acordo com um conjunto de variáveis sociais e estruturais que coocorrem com a variável dependente (cf. SANKOFF, 1988; NARO, 2004; GUY; ZILLES, 2007; TAGLIAMONTE, 2012). A fim de alcançar tal objetivo, desde a década de 1960, foram propostos alguns modelos matemáticos até que o modelo logístico fosse escolhido. A seguir apresentamos esses modelos.

Em 1969, William Labov propõe um modelo aditivo que pode ser representado como em (1):

$$f_t = f_o + f_1 + f_2 + \ldots \quad (1)$$

Em que f_o é a média global de aplicação da regra, um *input*, que serve como ponto de referência para o cálculo dos valores do efeito de cada fator componente de uma variável independente (f_i = f_1, f_2, ... f_n). O valor de cada f_i é a diferença entre a frequência média global de aplicação da regra e a frequência de aplicação no contexto que está sendo analisando, em outras palavras, f_i é o desvio, do valor verificado para cada fator, em relação ao *input* (f_o). Já f_t, é frequência geral de aplicação da regra, obtida pela soma de todos esses valores.

Esse modelo foi abandonado devido a problemas de natureza

técnica, uma vez que, como estamos falando de uma soma, não havia como garantir que o resultado dessa soma não fosse superior a 100% ou inferior a 0% (NARO, 2004, p. 20).

Em 1974 Henrietta Cedergren e David Sankoff propõem um modelo multiplicativo que, ao invés de frequências, utiliza cálculos de probabilidades gerando valores entre 0 e 1. Tal modelo é aplicável a uma ampla classe de regras e não possui as limitações técnicas do modelo aditivo, todavia, introduz uma complicação: o problema de precisar decidir se serão analisadas as probabilidades associadas à aplicação de uma regra variável ou à sua não aplicação (CEDERGREN; SANKOFF, 1974:337-338).

O modelo de aplicação é formalizado conforme (2):

$$p = p_o \times p_i \times p_j \ldots \quad (2)$$

Em que p é a probabilidade de que a regra seja aplicada, considerando todas as variáveis presentes no contexto. p_o é uma probabilidade *input*, comum a todos os ambientes e p_i é contribuição em probabilidade do fator *i*. Uma vez que, utilizamos o símbolo p para probabilidades de aplicação, $1 - p$ é probabilidade de que a regra não seja aplicada (CEDERGREN; SANKOFF, 1974:337). Por conseguinte, o modelo para análise da não aplicação da regra é formalizado como em (3):

$$(1-p) = (1-p_o) \times (1-p_i) \times (1-p_j) \ldots \quad (3)$$

E como demonstrado por Naro (2004) o modelo de aplicação tem um funcionamento satisfatório quando os fatores analisados desfavorecem a aplicação da regra e o modelo de não aplicação é apropriado para analisar fatores favorecedores. Destarte, a fim de resolver o problema supracitado, Rousseau e Sankoff (1978) aprimoram esse modelo chegando a modelagem de regressão

logística que trabalha com variáveis correlacionadas. Sankoff e Labov (1979) discorrem sobre suas vantagens afirmando que, diferente dos anteriores, o modelo logístico analisa de forma equilibrada, tanto fatores favorecedores, quanto desfavorecedores da aplicação de uma regra variável. Outrossim, os autores destacam, já naquele momento, o fato de que tal modelo já era amplamente utilizado na literatura, em várias áreas do conhecimento que utilizam a estatística e também em estudos linguísticos (cf. NARO; LEMLE, 1976; LEMLE; NARO, 1977; FASOLD, 1978; ROUSSEAU, 1978). Segundo Sankoff e Labov (1979), a função do modelo logístico, proposto por Rousseau e Sankoff (1978), é dada pela fórmula (4):

$$\log \left(\frac{P}{1 - P} \right) = \beta_0 + \beta_1 + \ldots + \beta_n$$

Nessa formalização *p* é substituído por *β*, conforme explicitado pelos próprios autores (Sankoff; Labov, 1979, p. 194). O modelo logístico será analisado, com maior riqueza de detalhes, na subseção 2.4.3.

Por meio do software Varbrul, esse modelo calcula, para cada fator, um *peso relativo (P.R.)* – um valor em probabilidade, numa escala entre 0 e 1 – que indica em que medida e em que direção cada fator afeta a taxa de aplicação da regra ou, em outras palavras, a probabilidade de aplicação da regra variável no contexto de cada fator (cf. TAGLIAMONTE, 2012). O valor do peso relativo deve ser interpretado da seguinte forma: um valor superior a 0,5 indica que o fator favorece a aplicação da regra, ao passo que um valor menor que 0,5 aponta que o fator a desfavorece. Um valor igual a 0,5, por sua vez, é um valor neutro e denota que o fator não tem efeito na aplicação da regra. Além disso, um valor muito próximo de 0 indica que a regra não se aplica no contexto daquele fator e um valor próximo de 1 aponta que a regra é categórica, ou seja, sempre

se aplica no contexto daquele fator (GUY; ZILLES, 2007, p. 41). Os pesos relativos, nos estudos incluídos nesta revisão sistemática, são reportados com o seguinte formato: P.R. .75, utilizando um "." (ponto) em vez de uma "," (vírgula) e omitindo o "0" (zero) que viria antes do ponto. Mas como demonstraremos, na seção 2.4.3, também é possível obter valores para o efeito dos fatores em outras unidades.

No início dos anos 2000 houve importantes desenvolvimentos nas técnicas estatísticas empregadas pela Sociolinguística Variacionista e o mais importante é a modelagem de efeitos mistos, ou modelo misto, que inclui, no modelo logístico, além das variáveis previsoras (independentes) fixas, variáveis aleatórias, que mudam a cada amostra, como *Informante* e *Item lexical* (cf. TAGLIAMONTE, 2012).

Variáveis aleatórias se diferenciam das variáveis fixas na medida em que estas possuem um número restrito de fatores e podem, facilmente, ser reproduzidas em outros estudos, em diferentes momentos e lugares, enquanto as aleatórias não podem. Os fatores das variáveis *Tonicidade da sílaba* (átona / tônica) e *Gênero do informante* (feminino / masculino), por exemplo, podem ser facilmente reproduzidos numa nova amostra de falantes, dado que, se repetirmos um mesmo estudo – como os realizados nos estudos primários desta revisão sistemática, que analisam a aplicação da regra variável da monotongação – a partir de uma nova amostra, provavelmente, teríamos palavras em que o ditongo ocorreria em sílabas átonas e tônicas e haveria homens e mulheres nessa amostra. De outra forma, essa nova amostra, dificilmente, conteria os mesmos informantes ou os mesmos itens lexicais da primeira amostra (JOHNSON, 2009; TAGLIAMONTE, 2012; OUSHIRO, 2017).

Não obstante, o modelo logístico, incluindo, ou não, variáveis aleatórias, não é o único que pode ser empregado para verificar o

efeito de variáveis previsoras (independentes) sobre uma variável resposta (dependente). Quando a variável resposta é numérica, como a altura de vogais, se medida pelos formantes em Hertz, por exemplo, o modelo adequado é o de regressão linear. Os trabalhos de Labov (1994) e Labov (2001) são exemplos de utilização desse modelo. Dessa forma, a escolha do modelo estatístico depende do tipo da variável resposta que se pretende estudar. Na seção 2.4, apresentamos algumas noções básicas de estatística e os modelos estatísticos de regressão, incluindo observações sobre modelos mistos.

2.4 Uma introdução à estatística e aos modelos de regressão

Para compreender os modelos de regressão é essencial conhecer algumas noções básicas de estatística sobre probabilidade, chance e razão de chances, além do teste de significância utilizado para testar hipóteses. Isto posto, nas subseções 2.4.1 e 2.4.2, apresentamos tais noções para introduzir, na subseção 2.4.3, os modelos de regressão, com destaque para o modelo de regressão logística.

2.4.1 Probabilidade, *odds* e *odds ratio*

Comecemos com noções sobre três medidas importantes para o entendimento do modelo de regressão logística: (i) probabilidade, (ii) chance ou *odds* e (iii) razão de chances ou *odds ratio (OR)*. Probabilidade é uma medida que indica a possibilidade de ocorrência de um evento qualquer, podendo ser obtida pela razão (divisão) entre o número de ocorrências do evento e o número total de ocorrências da amostra, sendo sempre um número entre 0 e 1.

Tabela 1: A monotongação de [ej] de acordo com a tonicidade da sílaba

Variantes	Átona	Tônica	Total
[ej]	149	1327	1476
[e]	161	544	705
Total	310	1871	2181

Fonte: Araújo (2000, adaptada)

Considerando a tabela 1 que mostra a distribuição da monotongação de [ej] de acordo com a tonicidade da sílaba, a probabilidade de ocorrer a monotongação de [ej] em qualquer tipo de sílaba é:

$$P_{total} = 705 / 2181 = 0,32$$

Já a probabilidade de ocorrer a monotongação numa sílaba tônica é obtida a partir da divisão do número de ocorrências de monotongação em sílabas tônicas pelo número total de ocorrências nesse contexto. Utilizando o mesmo raciocínio obtemos a probabilidade de que a monotongação de [ej] ocorra em sílaba átona:

$$P_{tônica} = 544 / 1871 = 0,29$$
$$P_{átona} = 161 / 310 = 0,52$$

A chance, ou *odds*, é a razão entre a probabilidade de que um evento ocorra e a probabilidade de que ele não ocorra. Sabendo que P é a probabilidade de um evento ocorrer (de sucesso), a probabilidade de o evento não ocorrer (ou de fracasso) é dada por 1 – P. Tal valor também pode ser obtido pela divisão do número de vezes que o ditongo foi mantido, nesse contexto, pelo total de ocorrências, no mesmo contexto. Dessa forma, a probabilidade de que a monotongação não ocorra numa sílaba tônica é:

$1 - P = 1 - 0{,}29 = 0{,}71 \quad = \quad 1327 / 1871 = 0{,}71$

Assim, a chance ou *odds* para cada uma das probabilidades calculadas acima é:

$Odds_{\text{total}} = P / 1 - P \rightarrow 0{,}32 / 1 - 0{,}32 = 0{,}47$
$Odds_{\text{tônica}} = P / 1 - P \rightarrow 0{,}29 / 1 - 0{,}29 = 0{,}41$
$Odds_{\text{átona}} = P / 1 - P \rightarrow 0{,}52 / 1 - 0{,}52 = 1{,}08$

A interpretação das chances é a seguinte: a probabilidade de que a monotongação ocorra numa sílaba tônica é 0,41 vezes a probabilidade de que o fenômeno não ocorra, isto é, 0,41 para 1. A razão de chances ou *odds ratio* compara as chances de ocorrência de um evento em dois diferentes contextos, ou fatores de uma variável, verificando o grau de associação entre fatores de uma mesma variável previsora (cf. OLIVEIRA, 2009; LEVSHINA, 2015):

$OR = Odds_{\text{átona}} / Odds_{\text{tônica}} \rightarrow 1{,}08 / 0{,}41 = 2{,}63$
$OR = Odds_{\text{tônica}} / Odds_{\text{átona}} \rightarrow 0{,}41 / 1{,}08 = 0{,}15$

A razão de chances entre sílaba átona e sílaba tônica mostra que a chance da monotongação ocorrer numa sílaba átona é 2,63 vezes maior do que a chance de que ocorra numa sílaba tônica. Já a chance da regra ser aplicada em sílaba tônica é 0,15 vezes a chance de acontecer numa sílaba átona.

2.4.2 Significância, hipótese nula e hipótese alternativa

Para identificar e explicar fenômenos linguísticos é preciso testar hipóteses e desenvolver modelos a partir dos quais seja possível

fazer previsões e a estatística inferencial fornece as ferramentas necessárias para tal tarefa. Uma dessas ferramentas é o teste de significância estatística que fornece como resultado o chamado *"valor-p"*, que pode ser entendido como a probabilidade de se observar determinada distribuição de dados em caso de a hipótese nula ser verdadeira. O que nos remete a dois importantes conceitos da estatística inferencial: *hipótese nula* (H_0) e *hipótese alternativa* (H_1). Esta última é a hipótese que está sendo testada, como, por exemplo, a afirmação de que há uma relação entre duas variáveis, enquanto a H_0, normalmente, é formulada como a negação da H_1, afirmando que não há relação entre as variáveis e que a distribuição dos dados observada resulta de uma flutuação aleatória e/ou erro de amostragem. O valor-p sempre é calculado, tendo como referência, a H_0, podendo ser entendido como a probabilidade de se observar determinado resultado, em caso de a hipótese nula ser verdadeira, contudo, uma definição mais ampla para o valor-p é: a probabilidade de que a distribuição dos dados tenha ocorrido ao acaso (cf. GUY; ZILLES, 2007; OUSHIRO, 2017).

Se essa probabilidade for muito baixa, a distribuição dos dados observada é estatisticamente significativa, indicando que a relação que está sendo testada é verdadeira, uma vez que a probabilidade dessa distribuição ocorrer por acaso é muito pequena. Portanto, quanto menor o valor-p, mais significativa é a distribuição dos dados analisada. Os valores do nível de significância podem ser obtidos por meio de vários testes estatísticos como, por exemplo, o Qui-quadrado que analisa, de forma comparativa, as proporções de duas variáveis nominais a fim de verificar se há associação entre essas variáveis (cf. GUY; ZILLES, 2007, LEVSHINA, 2015; OUSHIRO, 2017).

Convencionalmente, a comunidade científica costuma usar o limite de 0,05, ou 5%, para considerar algo como sendo muito pouco provável para acontecer ao acaso. Esse valor, denominado *nível*

α (alfa), é entendido como o limite, estabelecido pelo pesquisador, para rejeitar a hipótese nula: se o valor-p for igual ou superior a 0,05, a hipótese nula não pode ser rejeitada. No entanto, esse valor (0,05) é apenas uma convenção, logo, o pesquisador pode adotar outro valor para o nível α, maior ou menor, dependendo do objeto de estudo e do que se pretende fazer com o resultado obtido (GUY; ZILLES, 2007, p. 31-33; OUSHIRO, 2017).

O valor-p é uma medida de probabilidade, havendo sempre uma chance de erro. Por conseguinte, quando se estabelece um nível α de 5%, existe uma probabilidade de se chegar a conclusões equivocadas, em média, 5% das vezes, ou uma em cada vinte. Destarte, essa medida deve ser vista apenas como uma ferramenta para testar hipóteses e não como prova definitiva de que uma hipótese seja verdadeira ou falsa.

2.4.3 Modelos de Regressão

Antes de tratarmos dos modelos de regressão, revisemos a definição de modelo estatístico. Conforme Gries (2013, p. 253), modelo é uma caracterização formal da relação entre uma ou mais variáveis previsoras – e suas interações – e uma variável resposta. Isto posto, regressão é um tipo de modelo estatístico que ajuda a entender como o comportamento de uma ou mais variáveis pode mudar o comportamento da variável resposta. Essa relação entre variáveis pode ser analisada como um processo e neste os valores de $X_1, X_2, \ldots, Xn$ são chamados de variáveis previsoras ou explicativas e *Y* é chamado de variável resposta.

Desse modo, uma análise de regressão consiste no desenvolvimento de um modelo estatístico que possa ser utilizado para prever valores de uma variável resposta, com base nos valores de uma ou mais variáveis previsoras, dessa forma, seu objetivo é estimar e/ou prever o valor da variável resposta em função dos valores

conhecidos das variáveis previsoras (GUJARATI, 2000; LEVINE, BERENSON, STEPHAN, 2000). Mas uma análise de regressão pode ser usada com diferentes objetivos (cf. DIAS, 2005), tais como:

- Descrição: uma equação pode ser utilizada para resumir ou descrever um conjunto de dados em que a análise de regressão pode ser empregada para ajustar uma equação;
- Previsão: prever os valores da variável resposta;
- Estimação: estimar parâmetros desconhecidos de equações teóricas que representam o relacionamento entre as variáveis de interesse (cf. WERKEMA; AGUIAR, 1996).

As regressões podem ser simples ou múltiplas. Se o interesse é a relação de apenas uma variável previsora com a variável resposta temos um caso de regressão simples ou um modelo denominado univariado. Mas se o objetivo for relacionar a variável resposta a mais de uma variável previsora, a regressão é múltipla, já que a análise será multivariada (cf. DIAS, 2005; LEVSHINA, 2015; WINTER, 2020).

Em 1970, Nelder e Wedderburn propuseram os *Generalized Linear Models* (GLM) como uma extensão do modelo linear, por meio do qual só se pode modelar dados se a variável resposta é numérica contínua – que possa assumir qualquer valor numérico. Tais modelos permitem analisar outros tipos de variáveis respostas. Atualmente, a família GLM comporta vários modelos, entre eles, o próprio *Modelo Linear* que modela variáveis respostas numéricas contínuas; o *Modelo Poisson*, que modela variáveis respostas numéricas discretas, ou seja, contáveis que assumem um número finito de valores, como o número de alunos matriculados num determinado curso; o *Modelo Logístico*, que modela variáveis respostas binárias como sim / não, 0 / 1, ditongo / monotongo; o *Modelo Multinomial*, que modela variáveis resposta categóricas

que podem assumir mais de duas categorias nominais, como as possíveis realizações de fonema, como o rótico, em coda, que pode ser realizado como uma vibrante, um tepe, um retroflexo, entre outras (cf. PAULA, 2013). Logo, a escolha do modelo depende do tipo de variável resposta e, também, do tipo de distribuição dos dados. Assim, caso a variável resposta seja uma variável binária (tendo duas categorias), por exemplo, opta-se pelo o Modelo de Regressão Logística, mas se a variável resposta possui mais de duas categorias, utiliza-se o Modelo Multinomial. Apresentaremos, nas subseções 2.5.3.1, 2.5.3.2 e 2.5.3.3, noções introdutórias sobre o modelo linear e o modelo logístico da família GLM.

Quanto à distribuição dos dados é importante destacar que há vários tipos de distribuição – como: Normal, Binomial, Poisson, Exponencial, entre outras (cf. GRIES, 2013; PAULA, 2013; LEVSHINA, 2015), mas como trataremos, apenas, dos modelos linear e logístico, apresentaremos, somente, a distribuição Normal e a Binomial. A distribuição Normal é um tipo de distribuição contínua simétrica que quando representada num gráfico possui a forma de uma curva de sino, na qual medidas de tendências centrais como a média aritmética, a mediana (valor em relação ao qual metade de todas as observações é maior e metade é menor) e moda (valor da variável que ocorre com maior frequência) coincidem ou estão bem próximas. Dados modelados por uma regressão linear devem seguir esse tipo de distribuição (cf. GUY; ZILLES, 2007; GRIES, 2013; LEVSHINA, 2015; OUSHIRO, 2017; WINTER, 2020).

Uma distribuição binomial descreve situações em que os resultados de uma variável resposta estão agrupados em duas categorias, isto é, há apenas dois resultados possíveis, como sucesso ou falha. Tal tipo de distribuição caracteriza dados que podem ser modelados por uma regressão logística (cf. GRIES, 2013; LEVSHINA, 2015; OUSHIRO, 2017; WINTER, 2020). Com isso, vamos ao modelo de regressão linear.

2.4.3.1 Regressão linear simples

O modelo de regressão linear simples é utilizado quando se quer modelar a relação entre uma variável previsora, numérica ou categórica, e uma variável resposta numérica contínua (que pode assumir qualquer valor numérico).

Considerando duas variáveis X e Y. Dados n pares (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) , se Y é função[8] linear de X, pode-se estabelecer uma regressão linear simples, cujo modelo estatístico é dado pela fórmula 5:

$$Y_i = \beta_0 + \beta_1 X_i \qquad (5)$$

Em que:
i = 1, ..., n,
n é o tamanho da amostra.
Y_i é uma variável numérica contínua e representa o valor da variável resposta;
X representa o valor correspondente à variável previsora;
β_0 e β_1 são os parâmetros do modelo, que serão estimados, e que definem a reta de regressão. Mas antes de estimar os parâmetros é importante compreender o que cada um deles representa. O parâmetro β_0 é o coeficiente[9] linear, também chamado intercepto, e representa o ponto em que a reta linear corta o eixo y, em outras palavras é o valor de Y quando X é igual à zero. Já o parâmetro β_1 representa a inclinação da reta e, consequentemente, o efeito da variável previsora (X) sobre a variável resposta (Y). Esse parâmetro

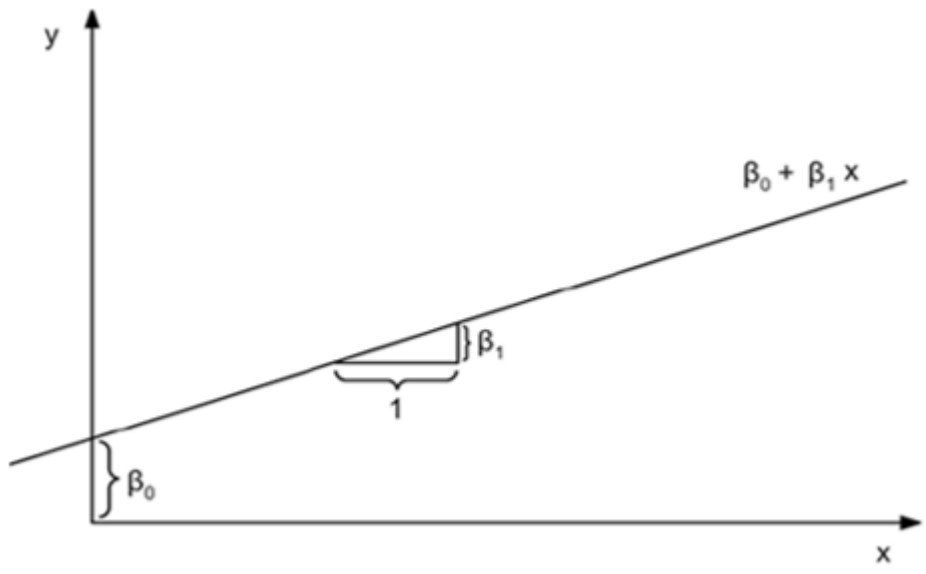
8 A função determina uma relação entre elementos de dois conjuntos. Podemos defini-la utilizando uma lei de formação, segundo a qual, para cada valor de x, temos um valor de f(x). A formalização matemática para a definição de função é dada por: *Seja X um conjunto com elementos de x e Y um conjunto dos elementos de y, temos que: f:* $x \rightarrow y$. Chamamos x de domínio e f(x) ou y de imagem da função (IEZZI et al., 2011).

9 Coeficiente é um número que, colocado à esquerda de uma quantidade algébrica, lhe serve de fator multiplicativo (IEZZI et al., 2011).

é chamado coeficiente de regressão ou coeficiente angular. Para cada aumento de uma unidade na variável X, o valor Y aumenta β_1 unidades, isto é, β_1 mostra o quanto o Y aumenta, ou diminui, a cada unidade de aumento de X (GUIMARÃES, 2012).

A relação entre X e Y é linear e os valores de X são fixos (ou controlados), isto é, X não é uma variável aleatória. Num gráfico de dispersão essa relação resulta numa reta. A interpretação geométrica dos parâmetros β_0 e β_1 está representada na figura 1. Porém, nem todos os pares (X e Y) de valores observados que geram os pontos no gráfico, a partir dos quais a reta é traçada, coincidem exatamente com a reta. Alguns desses valores observados podem não ser previstos pelo modelo podendo, por conseguinte, se distanciar da reta traçada. Essas diferenças são chamadas de *resíduos*, os quais podem, dessa forma, ser definidos como a diferença entre um valor previsto, ou estimado, pelo modelo e um valor observado na distribuição dos dados (GUIMARÃES, 2012; OUSHIRO, 2017).

Figura 1: Representação de uma reta regressora



Fonte: Portal Action (2020)

Os valores de β_0 e β_1 podem ser estimados através do Método dos Mínimos Quadrados (MMQ), no qual β_0 e β_1 são obtidos de forma que a soma dos quadrados das diferenças entre os valores

observados de Y e os valores obtidos a partir da reta ajustada para os mesmos valores de X é mínima (cf. TOLEDO; OVALLE, 1985; DIAS, 2005; GUIMARÃES, 2012).

2.4.3.2 Regressão linear múltipla

A regressão linear múltipla modela relações entre mais de uma variável previsora e uma variável resposta numérica contínua (cf. WINTER, 2020). Esse modelo é dado pela fórmula 6:

$$Y_i = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \quad (6)$$

Em que *Y* representa a variável resposta, X_1, $X_{2,}$,...,X_n representam os valores ou níveis das variáveis previsoras. Como no modelo linear, o parâmetro β_o é o coeficiente linear, também chamado intercepto, e representa o ponto em que a reta linear corta o eixo y, o valor de Y quando X é igual à zero. Os parâmetros β_1, β_2,..., β_n representam os coeficientes de regressão ou coeficientes angulares correspondendo aos valores que multiplicarão as variáveis previsoras representadas por X_1, X_2 ..., X_n, ou seus efeitos na variável resposta. O parâmetro β_1 indica uma mudança na resposta média a cada unidade de mudança em X_1, quando as demais variáveis são mantidas fixas. De forma semelhante β_2 indica uma mudança na resposta média a cada unidade de mudança em X_2, quando as demais variáveis são mantidas constantes (GUIMARÃES, 2012).

Os parâmetros também podem ser estimados pelo Método dos Mínimos Quadrados (GUIMARÃES, 2012). Os valores desses parâmetros também podem ser obtidos através da utilização do software Rbrul ou R. Com isso, chegamos ao modelo de regressão logística.

2.4.3.3 Regressão Logística

No modelo de regressão logística a variável resposta Y é binária , isto é, uma variável que pode assumir um de dois valores ou categorias possíveis, como por exemplo, como por exemplo, Y = 0 e Y = 1 que podem ser denominados “fracasso” e “sucesso”, respectivamente. Sucesso é o evento de interesse, como a aplicação da regra variável de monotongação (redução do ditongo a uma vogal simples ou monotongo), e fracasso é a não ocorrência do evento de interesse, para usar o mesmo exemplo, a não aplicação da regra de monotongação (manutenção do ditongo). As variáveis previsoras, por sua vez, podem ser numéricas ou categóricas. A regressão logística permite que se estime o logaritmo da chance ou a probabilidade de ocorrência de determinado evento considerando uma ou mais variáveis previsoras. Mas o interesse desse tipo de análise é trabalhar com múltiplas variáveis previsoras, verificando o efeito simultâneo dessas variáveis, com a finalidade de se chegar a um modelo para descrever, explicar e prever o comportamento da variável resposta (cf. GUY; ZILLES, 2007; LEVSHINA, 2015; OUSHIRO, 2017).

Podemos dizer que o modelo de regressão linear e logística tem muito em comum (GUIMARÃES, 2012; PAULA, 2013). Partindo do modelo linear, em (7), para chegarmos ao logístico precisamos transformar as variáveis previsoras lineares de forma que a variável resposta seja dada em probabilidade (0, 1) utilizando-se uma função em (8). Dessa forma, a representação gráfica da relação entre a variável resposta e uma previsora deixa de ser uma reta e passa a ser uma curva com o formato da letra “S”.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_{n.} \quad (7)$$

$$P\,(y_i = 1) = f\,(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n) \quad (8)$$

Desse modo, a probabilidade de sucesso de um evento é dada em função das variáveis previsoras. No entanto, há duas condições as quais essa função precisa atender:

P deve:
(i) ser maior ou igual a zero (sempre positivo) e
(ii) menor ou igual a 1:

De forma sintética: $0 \leq P \leq 1$. Para que P seja sempre positivo usamos o exponencial[10] das variáveis previsoras (9) e para que P seja menor do que 1 aplicamos a divisão em (10), já que a razão entre qualquer valor e 1 mais ele mesmo é sempre menor que 1.

$$e^{\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn} > 0 \qquad (9)$$

$$0 \leq \frac{e^{\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn}}{1 + e^{\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn}} \leq 1 \qquad (10)$$

Assim, a função logística pode ser reescrita como em (11):

$$P(y_i = 1) = \frac{1}{1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn)}} \qquad (11)$$

A fim de isolar as variáveis previsoras, utilizamos a função *logit*, baseada no logaritmo natural (*ln*) [11] da chance, uma operação inversa à exponenciação, e obtemos a função 12.

10 A exponenciação é uma operação que consiste em transformar um valor de base em expoente, nesse caso, exponente de e, que é o número de Euler ou número neperiano, um número irracional igual a 2,718281828459... (IEZZI et al., 2011)

11 Logaritmo é o expoente da base e pode ser formulado da seguinte forma: sejam a e b dois números reais positivos ($a \neq 1$, $b > 0$ e $a > 0$), denomina-se logaritmo de a na base b o expoente x ($\log_a b = x$), sendo $b^x = a$. O logaritmo natural, representado por ***ln***, é um logaritmo de base ***e*** (o número de Euler ou número neperiano), um número irracional igual a 2,718281828459... (IEZZI et al., 2011)

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n = logit\,(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n) \quad (12)$$

Em que $\frac{P}{1-P}$ é a chance de um evento ocorrer em relação à chance de que o mesmo evento não ocorra ou *odds*, como visto na seção 2.4.1, e $\ln\left(\frac{P}{1-P}\right)$ é a variável resposta, dada em *logit*, que pode ser lida como o logaritmo natural[12] de odds.

$\beta_0, \beta_1, \ldots, \beta_n$ são os parâmetros do modelo e $X_0, X_1, \ldots, X_n$ são as variáveis previsoras. Sendo que, como na regressão linear, β_0 é o coeficiente linear, também chamado intercepto, e corresponde ao valor da variável resposta quando os X são todos igual a zero. Os parâmetros $\beta_1, \beta_2, \ldots, \beta_n$ representam os coeficientes angulares, correspondendo aos valores que multiplicarão as variáveis previsoras ou níveis (fatores) de uma variável previsora categórica. Assim, esses coeficientes correspondem ao efeito de cada fator de uma variável previsora categórica ou da própria variável, quando esta for numérica, sobre a variável resposta (PAULA, 2013).

No modelo de regressão logística, a variável resposta é dada em *logodds* (logaritmo de odds), assim como o intercepto. Já os coeficientes angulares, apesar de também serem fornecidos em *logodds*, são obtidos, com base no logaritmo de *odds ratio*. Valores em *logodds* podem assumir qualquer valor, negativo ou positivo. Sua interpretação se dá da seguinte forma: (i) valores negativos representam variáveis ou fatores que desfavorecem a aplicação da regra variável; (ii) valores positivos correspondem a fatores que favorecem a regra; (iii) um valor igual a zero indica que o fator

12 Entende-se $\boldsymbol{ln\ x = b}$ como: logaritmo natural de $\boldsymbol{x}$ é igual a $\boldsymbol{b}$. Sabendo que o logaritmo é o expoente e que a base de um logaritmo natural é o número de Euler ($\boldsymbol{e}$) (IEZZI *et al.*, 2011), esse logaritmo corresponde a seguinte equação exponencial: $\boldsymbol{e^{\,b} = x}$ que se lê: o número de Euler elevado ao expoente $\boldsymbol{b}$ é igual a $\boldsymbol{x}$ ou exponencial de $\boldsymbol{b}$ é igual a $\boldsymbol{x}$.

ou variável não tem efeito na regra (cf. GRIES, 2013, OUSHIRO, 2017; WINTER, 2020). Entretanto, se quisermos obter resultados em probabilidade, numa escala entre 0 e 1, para os coeficientes das variáveis previsoras, é possível retornar para a função (11), que gera valores probabilísticos, aplicando a função inversa do *logit* em (13):

$$\frac{1}{1+ e^{x}} \qquad (13)$$

Que permite que a função logística seja reescrita seguindo os passos em (14), (15), (16) e (17):

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \qquad (14)$$

$$\frac{P}{1-P} = e^{\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn} \qquad (15)$$

$$P = \frac{e^{\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn}}{1+ e^{\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn}} \qquad (16)$$

$$P = \frac{1}{1+ e^{-(\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn)}} \qquad (17)$$

Dessa forma, obtemos, novamente, a função logística que gera valores probabilísticos (18):

$$\rightarrow \quad P(Yi = 1) = \frac{1}{1+ e^{-(\beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn)}} \qquad (18)$$

Essa função pode ser lida como: a probabilidade de sucesso

(ou de que a regra variável seja aplicada) é igual a razão entre 1 e 1 acrescido do exponencial de β_0 mais $\beta_1 X_1$ mais $\beta_2 X_2$ e assim, sucessivamente, a depender do número de variáveis previsoras consideradas. Os valores em *logodds* e em probabilidades fornecem a mesma informação, mas de formas diferentes (cf. OUSHIRO, 2017; JOHNSON, 2009; WINTER, 2020). A tabela 2 mostra a relação entre valores em probabilidade e em *logodds*.

Tabela 2: Relação entre as medidas de probabilidade e *logodds*

Probabilidade	***Logodds***
.00	- ∞
.10	-2,2
.20	-1,39
.30	-0,85
.40	-0,4
.50	0
.60	0,4
.70	0,85
.80	1,39
.90	2,2
1.0	+ ∞

Fonte: Johnson (2009, p. 361, adaptada)

No modelo logístico, os valores dos parâmetros ou coeficientes são calculados pelo método da máxima verossimilhança que consiste em estimar os parâmetros do modelo utilizando as estimativas que tornam máximo o valor da função de verossimilhança, ou seja, entre todos os valores possíveis, esse método encontra os valores mais prováveis de terem gerado os dados observados (cf. TAGLIAMONTE, 2006; OLIVEIRA, 2009). Tais valores podem ser obtidos por meio da utilização de softwares ou programas computacionais como o pacote de programas Varbrul (ou alguma

de suas versões), o SPSS e, mais recentemente, o Rbrul e o R, que automatizam os cálculos (cf. JOHNSON, 2009; OLIVEIRA, 2009; TAGLIAMONTE, 2012).

Contudo, para que os resultados obtidos sejam confiáveis é necessário verificar se o modelo de regressão logística não viola alguns pressupostos básicos: (i) não pode haver multicolinearidade entre as variáveis previsoras, que ocorre quando algumas dessas variáveis incluídas no modelo se referem a um mesmo efeito, ainda que de uma forma diferente; (ii) verificar se o efeito de cada uma das variáveis é independente. Quando isso não ocorre, há interação – isto é, uma relação de dependência – entre variáveis, o que deve ser considerado no modelo. Uma das formas de identificar uma interação entre variáveis é examinar a relação entre os valores dos efeitos (pesos relativos, se estes forem calculados em probabilidade) e os valores percentuais de aplicação da regra variável, calculados para cada fator de uma variável. Quando há interação, esses valores ficam desalinhados, isto é, o modelo estatístico fornece valores de efeitos que não correspondem aos percentuais de aplicação da regra. Softwares como o Rbrul e o R possibilitam que as interações sejam identificadas de forma bastante simples e rápida, como veremos na seção 2.5; (iii) a distribuição dos resíduos deve ser normal, com valores simétricos entre os quartis e uma mediana próxima a zero; e (iv) é preciso verificar se cada observação é independente das outras (cf. TAGLIAMONTE, 2012; OUSHIRO, 2017).

2.4.3.4 Dados linguísticos e modelos mistos

Modelos de regressão simples pressupõem uma independência entre cada dado coletado da população – que compõe a amostra de dados. Todavia, em amostras de dados linguísticos, isso raramente acontece, uma vez que se trabalha com um número pequeno de informantes, ou seja, os dados vêm de um pequeno conjunto da

população e de cada informante, é extraído um determinado número de dados, logo, os dados de um informante não são independentes uns dos outros, ademais, os itens lexicais também se repetem nos dados (cf. JOHNSON, 2009; TAGLIAMONTE, 2012; OUSHIRO, 2017; LIMA Jr.; GARCIA, 2021).

Assim, boa parte da variabilidade nos dados se deve ao informante, dado que cada um deles "traz aos dados uma variação intrínseca e individual" (LIMA Jr.; GARCIA, 2021, p. 13) e, da mesma forma, cada item lexical que aparece diversas vezes nos dados pode exercer diferentes efeitos sobre a variável resposta. Desse modo, tanto o informante quanto o item lexical são variáveis previsoras, mas o efeito dessas variáveis é aleatório, já que alterando os participantes e/ou os itens lexicais poderíamos constatar resultados diferentes e, se repetíssemos um mesmo estudo seria muito difícil obter uma amostra com os mesmos informantes e os mesmos itens lexicais, o que caracteriza essas variáveis como sendo aleatórias (OUSHIRO, 2017; LIMA Jr.; GARCIA, 2021).

O objetivo de um modelo de regressão não é alcançar resultados válidos apenas para a amostra analisada, mas poder generalizar seus resultados para a população. Dessa forma, é importante informar ao modelo estatístico a existência de efeitos de variáveis aleatórias para que os resultados dos efeitos das variáveis fixas – os efeitos que interessam no estudo – sejam ajustados e, para tanto, é preciso empregar um modelo de efeitos mistos. Esse tipo de modelo é capaz de considerar os efeitos de variáveis aleatórias como o *Informante* e o *Item lexical*, caracterizando-se, portanto, como um modelo mais robusto, já que leva em consideração a não independência dos dados em estudos linguísticos (cf. JOHNSON, 2009; TAGLIAMONTE, 2012; OUSHIRO, 2017; LIMA Jr.; GARCIA, 2021). Isto posto, vamos aos softwares que podem ser utilizados para realizar os cálculos estatísticos.

2.5 Os softwares Varbrul, Rbrul e R

O software Varbrul (*Variable Rule Program*) foi criado, no início dos anos 1960, especialmente para realizar a modelagem estatística de uma regra variável em estudos que utilizavam os pressupostos da Sociolinguística Variacionista. Contudo, desde o início dos anos 2000, passou a haver um ceticismo quanto ao fato dessa ferramenta continuar a ser a mais apropriada para realizar os cálculos estatísticos em estudos variacionistas (JOHNSON, 2009; TAGLIAMONTE, 2012).

Em 2009, durante o 38º *New Ways of Analyzing Variation* (NWAV), evento anual, realizado na Universidade de Ottawa, começa uma discussão a respeito de limitações do Varbrul, em função de novas teorias linguísticas que tratam do papel da frequência do item lexical e/ou consideram a variabilidade nos dados que pode ser atribuída ao indivíduo (cf. BAYLEY, 2004; JOHNSON, 2009; OLIVEIRA, 2012). Tal discussão ocorreu no decorrer de um workshop intitulado *Using statistical tools to explain linguistic variation - A state of the art workshop for NWAV 38*, coordenado pela professora Sali Tagliamonte e continuou em 2010, na 39ª versão do mesmo evento, na Universidade do Texas, durante dois workshops, o primeiro ministrado pelo professor Daniel Ezra, nomeadamente, *Quantitative Analysis with Rbrul and R* e o segundo coordenado pela professora Sali Tagliamonte. No 41º NWAV, na Universidade de Indiana, em 2012, houve outro workshop, sobre o mesmo tema, ministrado pelo professor John Paolillo: *Linguistic Variation, Theory-building and Statistics: Toward an Integrated Perspective* (SCHERRE, 2012, p. 6-7). Tais discussões apontaram para a existência de novas ferramentas (softwares) que não possuem as limitações do Varbrul, sobretudo, quanto à inclusão, na análise estatística, das variáveis aleatórias *Item lexical* e *Informante*.

Para considerar o efeito de variáveis aleatórias na análise de

uma regra variável é necessário empregar um modelo misto, e as versões do Varbrul modelam apenas variáveis previsoras categóricas e de efeitos fixos, não estando equipado para trabalhar com efeitos mistos. Ademais, o Varbrul executa apenas a modelagem estatística de regressão logística e, por conseguinte, só pode analisar variáveis respostas (dependentes) binárias (cf. GUY; ZILLES, 2007; JOHNSON, 2009; TAGLIAMONTE, 2012).

Na verdade, a modelagem utilizada no Varbrul – regressão logística – está disponível em qualquer software que executa cálculos estatísticos, no entanto, o Varbrul apresenta resultados em um formato raramente visto em outros programas e usa uma terminologia diferente dos demais softwares: as variáveis previsoras são denominadas *grupos de fatores* e o que a estatística denomina *níveis* é chamado de *fatores* (JOHNSON, 2009, p. 360). Nas subseções 2.5.1, 2.5.2 e 2.5.3, discutiremos a utilização e o *output* do software Varbrul – utilizado em todos os estudos incluídos nesta revisão sistemática – e de dois outros softwares que podem ser utilizados para realizar, além de uma regressão logística, outras modelagens estatísticas, inclusive modelos mistos, a saber, o Rbrul e o R, por meio de sua interface RStudio. A fim de demonstrar a utilização do Rbrul e do R, nas subseções 2.5.2 e 2.5.3, respectivamente, analisamos a aplicação da regra variável de apagamento do rótico em coda no português santomense, num *corpus* obtido a partir de 12 entrevistas de fala espontânea com duração de, aproximadamente, 60 minutos, realizadas em 2016 e 2019 em São Tomé e Príncipe (BALDUINO, 2016; 2019). O software Varbrul (*Variable Rule Program*) foi criado, no início dos anos 1960, especialmente para realizar a modelagem estatística de uma regra variável em estudos que utilizavam os pressupostos da Sociolinguística Variacionista. Contudo, desde o início dos anos 2000, passou a haver um ceticismo quanto ao fato dessa ferramenta continuar a ser a mais apropriada para realizar os cálculos estatísticos em estudos variacionistas (JOHNSON, 2009;

TAGLIAMONTE, 2012).

Em 2009, durante o 38º *New Ways of Analyzing Variation* (NWAV), evento anual, realizado na Universidade de Ottawa, começa uma discussão a respeito de limitações do Varbrul, em função de novas teorias linguísticas que tratam do papel da frequência do item lexical e/ou consideram a variabilidade nos dados que pode ser atribuída ao indivíduo (cf. BAYLEY, 2004; JOHNSON, 2009; OLIVEIRA, 2012). Tal discussão ocorreu no decorrer de um workshop intitulado *Using statistical tools to explain linguistic variation - A state of the art workshop for NWAV 38*, coordenado pela professora Sali Tagliamonte e continuou em 2010, na 39ª versão do mesmo evento, na Universidade do Texas, durante dois workshops, o primeiro ministrado pelo professor Daniel Ezra, nomeadamente, *Quantitative Analysis with Rbrul and R* e o segundo coordenado pela professora Sali Tagliamonte. No 41º NWAV, na Universidade de Indiana, em 2012, houve outro workshop, sobre o mesmo tema, ministrado pelo professor John Paolillo: *Linguistic Variation, Theory-building and Statistics: Toward an Integrated Perspective* (SCHERRE, 2012, p. 6-7). Tais discussões apontaram para a existência de novas ferramentas (softwares) que não possuem as limitações do Varbrul, sobretudo, quanto à inclusão, na análise estatística, das variáveis aleatórias *Item lexical* e *Informante*.

Para considerar o efeito de variáveis aleatórias na análise de uma regra variável é necessário empregar um modelo misto, e as versões do Varbrul modelam apenas variáveis previsoras categóricas e de efeitos fixos, não estando equipado para trabalhar com efeitos mistos. Ademais, o Varbrul executa apenas a modelagem estatística de regressão logística e, por conseguinte, só pode analisar variáveis respostas (dependentes) binárias (cf. GUY; ZILLES, 2007; JOHNSON, 2009; TAGLIAMONTE, 2012).

Na verdade, a modelagem utilizada no Varbrul – regressão logística – está disponível em qualquer software que executa

cálculos estatísticos, no entanto, o Varbrul apresenta resultados em um formato raramente visto em outros programas e usa uma terminologia diferente dos demais softwares: as variáveis previsoras são denominadas *grupos de fatores* e o que a estatística denomina *níveis* é chamado de *fatores* (JOHNSON, 2009, p. 360). Nas subseções 2.5.1, 2.5.2 e 2.5.3, discutiremos a utilização e o *output* do software Varbrul – utilizado em todos os estudos incluídos nesta revisão sistemática – e de dois outros softwares que podem ser utilizados para realizar, além de uma regressão logística, outras modelagens estatísticas, inclusive modelos mistos, a saber, o Rbrul e o R, por meio de sua interface RStudio. A fim de demonstrar a utilização do Rbrul e do R, nas subseções 2.5.2 e 2.5.3, respectivamente, analisamos a aplicação da regra variável de apagamento do rótico em coda no português santomense, num *corpus* obtido a partir de 12 entrevistas de fala espontânea com duração de, aproximadamente, 60 minutos, realizadas em 2016 e 2019 em São Tomé e Príncipe (BALDUINO, 2016; 2019).

2.5.1 O Varbrul

O software Varbrul é um pacote de programas que foi criado especialmente para conduzir análises estatísticas em estudos sociolinguísticos, especificamente, para realizar análises de regras variáveis que controlam variáveis linguísticas binárias (com duas realizações possíveis), por meio de uma regressão logística, sendo esta a única modelagem estatística que o software realiza (cf. TAGLIAMONTE, 2006; GUY; ZILES, 2007, TAGLIAMONTE, 2012). A utilização dessa ferramenta não exige muito conhecimento na área de estatística e fornece resultados, num formato, com o qual, os sociolinguistas estão habituados a trabalhar.

O software foi desenvolvido pelo matemático David Sankoff na década de 1970 e aprimorado nos anos seguintes (cf. OLIVEIRA,

2009), quando foram criadas novas versões dessa ferramenta:

- Varbrul 2S (SANKOFF, 1972);
- MacVarb (GUY; LIPA, 1987);
- Varbrul 3M (ROUSSEAU, 1978);
- PC-VARB (PINTZUK; SANKOFF, 1982);
- Goldvarb 2.0 (RAND; SANKOFF, 1990);
- Goldvarb 2.1 (RAND; SANKOFF, 1992);
- Goldvarb 2001 (LAWRENCE; TAGLIAMONTE, 2001);
- R-VARB (PAOLILLO, 2002);
- Goldvarb X (SANKOFF; TAGLIAMONTE; SMITH, 2005).

As versões do Varbrul analisam apenas variáveis previsoras categóricas, por conseguinte, é necessário codificar variáveis numéricas, como a idade dos informantes, por exemplo, em categorias, geralmente, faixas etárias (1ª, 2ª e 3ª, em que cada faixa é constituída por um intervalo de idade como tais: como *menores de 20*, *entre 20 e 40*, e *acima de 40*). Além disso, o software requer que a base de dados seja um arquivo codificado, de forma específica, de modo que, cada fator de uma variável seja representado por um único caractere (cf. JOHNSON, 2009), tornando o trabalho do pesquisador mais oneroso, o que, como veremos nas próximas seções deste capítulo, não é necessário quando se utiliza outros softwares como o Rbrul e o R.

Antes de realizar a modelagem dos dados, o software fornece uma visão geral da distribuição dos dados, por meio de tabelas de frequência e percentuais, quando podem ser detectados os termos: *knockouts* e *singletons*. O termo *knockout*, que pode aparecer num *output* do Goldvarb, indica que um fator corresponde, num determinado momento da análise, à frequência de uma das variantes da variável dependente, de 0% ou 100%, destarte, tal fator não pode ser considerado na análise de uma regra variável, no Varbrul, posto

que, no contexto desse fator, não houve variação. Na maioria dos casos, fatores, nos quais ocorre *knockout*, são removidos da análise (cf. TAGLIAMONTE, 2006; GUY; ZILLES, 2007). Já *singleton* significa que há apenas um fator num grupo de fatores. Esse grupo com apenas um fator pode ser removido da análise ou dividido em outras categorias (cf. TAGLIAMONTE, 2006). Após essa verificação inicial dos dados, pode-se executar a análise da regra variável propriamente dita.

O Goldvarb X, última versão do Varbrul, oferece duas formas para conduzir a análise de uma regra variável: (i) *binomial one-step* e (ii) *binomial step up/step-down*. A primeira analisa todas as variáveis – ou grupos de fatores – ao mesmo tempo e a segunda realiza uma análise nivelada, com computações em um *step* por vez. A maioria dos estudos emprega esta última (TAGLIAMONTE, 2006, p. 139-140).

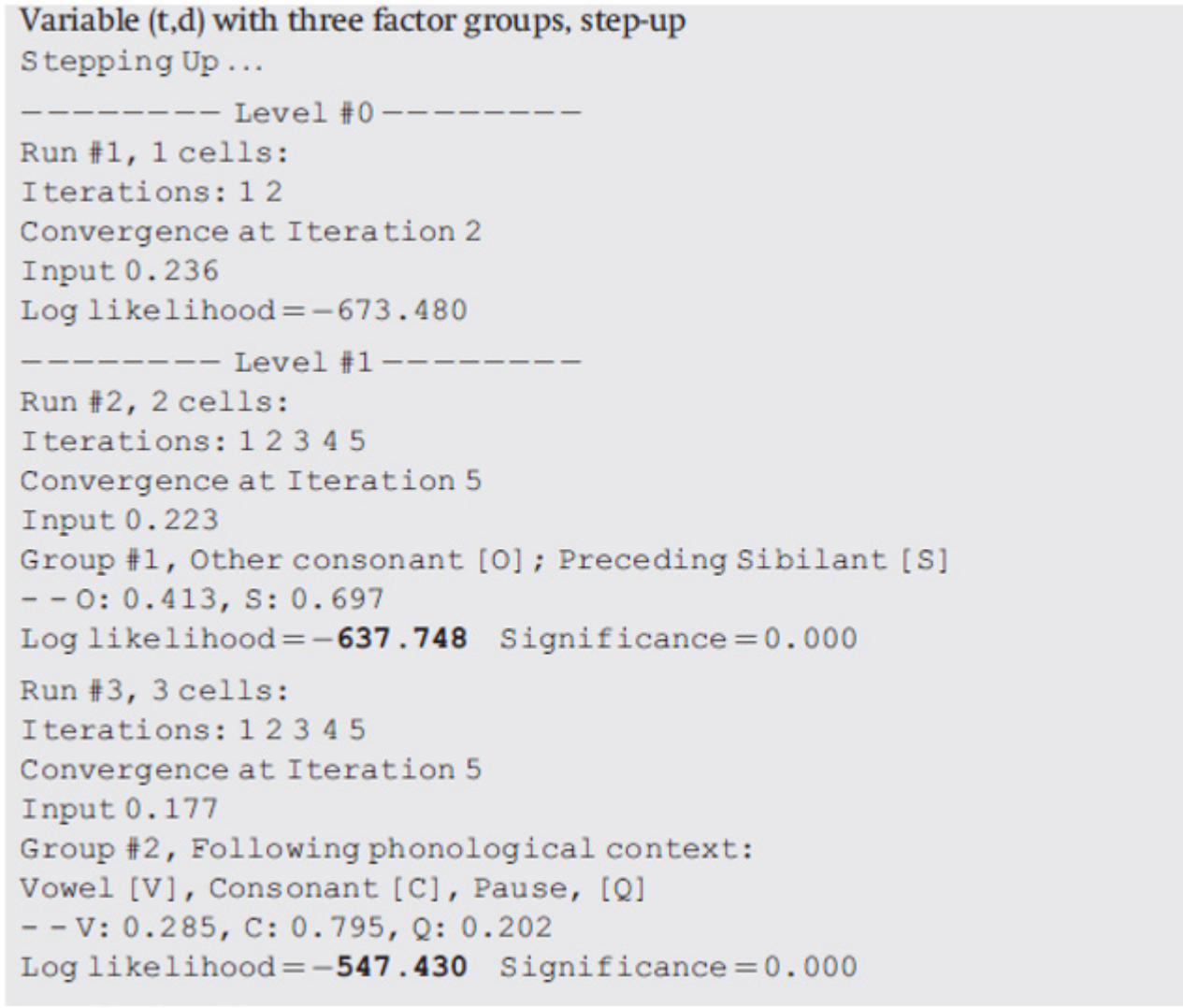
Quando uma regressão logística é executada empregando o método *binomial step-up/step-down,* no Goldvarb X, é apresentado o *step-up* e, em seguida, o *step-down*. O primeiro passo do *step-up* é encontrar a variável que causa a mudança mais significativa no modelo testando todas, a fim de, determinar qual delas melhora o *likelihood* de forma mais significativa. Então, o programa mantendo a variável mais significante tenta adicionar a segunda variável que melhora o *likelihood* significativamente. Esse processo continua até que não haja mais variáveis que possam ser adicionadas gerando uma melhoria estatisticamente significante. A análise é apresentada em *levels* (cf. figura 2) com as chamadas "rodadas" (*Run*), para cada uma das quais, é apresentado um número de "iterações", o ponto de convergência e o *log-likelihood* (cf. TALIAMONTE, 2006), termos que explicitamos a seguir.

As iterações são ciclos de ajustes nos valores dos pesos relativos dos fatores, que procuram um resultado otimizado, entre o modelo estatístico e os dados observados. Assim, as iterações

informam o número de ciclos realizados, a fim de se obter tal resultado. O ponto de convergência ocorre, justamente, quando o resultado otimizado é obtido. Já o *log-likelihood* ou logaritmo da verossimilhança é um valor numérico que mede a qualidade da aproximação entre o modelo e os valores observados. Quanto mais próximo de zero é esse valor, maior é a qualidade dessa aproximação. O valor absoluto do *log-likelihood* também varia em função da quantidade de dados. Logo, valores absolutos *log-likelihood* só podem ser comparados quando se mantém o mesmo número de dados na análise (cf. GUY; ZILLES, 2007).

A análise começa, no *nível 0,* com um modelo que não inclui nenhuma variável e um *input* global – uma medida global da taxa de aplicação da regra ou probabilidade geral de aplicação da regra – e depois vai acrescentando as variáveis previsoras uma a uma (cf. TALIAMONTE, 2006).

Figura 2: Output do step-up fornecido pelo Goldvarb X



```
Variable (t,d) with three factor groups, step-up
Stepping Up...
-------- Level #0 --------
Run #1, 1 cells:
Iterations: 1 2
Convergence at Iteration 2
Input 0.236
Log likelihood = -673.480
-------- Level #1 --------
Run #2, 2 cells:
Iterations: 1 2 3 4 5
Convergence at Iteration 5
Input 0.223
Group #1, Other consonant [O]; Preceding Sibilant [S]
- - O: 0.413, S: 0.697
Log likelihood = -637.748  Significance = 0.000
Run #3, 3 cells:
Iterations: 1 2 3 4 5
Convergence at Iteration 5
Input 0.177
Group #2, Following phonological context:
Vowel [V], Consonant [C], Pause, [Q]
- - V: 0.285, C: 0.795, Q: 0.202
Log likelihood = -547.430  Significance = 0.000
```

Fonte: Tagliamonte (2006, p. 141)

A figura 2 mostra a primeira parte de um *output* de uma análise *step-up*, do apagamento de /t/ e /d/ em posição final, considerando três variáveis previsoras, aqui, denominadas, *grupos de fatores*. Nesse exemplo, *no nível 1* são realizadas três rodadas, cada uma verificando a significância de um dos três grupos de fatores. É selecionado o grupo de fatores que está na rodada com o valor de *log-likelihood* mais próximo de zero. No *nível 3*, será mantido o primeiro grupo de fatores nas rodadas e, em cada uma, será adicionada um segundo grupo de fatores que mais aproximar o *log-likelihood* de zero. Sendo que, só são selecionadas variáveis com um nível de significância menor que 0,05 (cf. TALIAMONTE, 2006). Esse é o

chamado *valor-p*, sobre o qual discorremos na subseção 2.4.2.

A ordem de seleção das variáveis forma uma organização hierárquica, ordenando essas variáveis de acordo com sua força ou importância, para a aplicação da regra variável. Mas, também, é possível realizar a organização hierárquica das variáveis previsoras, a partir do valor do range entre os pesos reativos dos fatores de cada variável, ou seja, a diferença entre o maior e o menor peso relativo dos fatores de uma mesma variável. O *range* é, assim, uma medida não estatística que indica a força de uma variável, quanto maior for esse valor, maior será a força da variável (TAGLIAMONTE, 2012, p. 123-127).

O *step-down* é baseado no mesmo princípio que o *step-up*, no entanto ocorre na direção contrária: o programa começa calculando o *likelihood* do modelo completo, com todas as variáveis incluídas na regressão e vai descartando as variáveis cuja exclusão reduz, menos significativamente, o *likelihood*, o que é feito utilizando o teste de Qui-quadrado (TAGLIAMONTE, 2006, p. 140-143).

Os valores dos efeitos dos fatores, das variáveis selecionadas, ou parâmetros do modelo, reportados em probabilidade, numa escala entre 0 e 1, são denominados *pesos relativos* (cf. TAGLIAMONTE, 2006; GUY; ZILLES, 2007) e o método empregado para realizar os cálculos é denominado *sum contrasts* (JOHNSON, 2009, p. 361), ou desvio da média.

Comumente, numa regressão logística, os fatores de uma variável previsora categórica são codificados de forma que um nível ou fator é escolhido como referência ou *baseline*. Esse tipo de codificação é utilizado por softwares que executam cálculos estatísticos, como o SPSS (cf. OLIVEIRA, 2009, p. 106) e o R (cf. GRIES, 2013; LEVSHINA, 2015; OUSHIRO, 2017; WINTER, 2020), e a partir dele, os valores dos efeitos dos fatores de uma variável são calculados em relação ao fator ou nível de referência dessa mesma variável. Desse modo, o valor do parâmetro, ou coeficiente, de um

nível é a estimativa do efeito da troca do nível de referência para o nível em questão (JOHNSON, 2009, p. 361). Considerando os dados da tabela 1, se aplicássemos um modelo de regressão logística para verificar apenas o efeito dos níveis, ou fatores, da variável *Tonicidade* sobre a aplicação da regra variável de monotongação do ditongo [ej], por exemplo, o efeito do fator *tônica* seria calculado em relação ao fator de referência, que nesse caso é *átona*:

$$OR = Odds_{\text{tônica}} / Odds_{\text{átona}} \rightarrow 0{,}41 / 1{,}08 = 0{,}15$$

Esse valor está em *odds ratio* e é o mesmo calculado na seção 2.4.1, podendo ser reportado nessa unidade de medida. Todavia, o modelo exemplificado contém apenas uma variável previsora. Numa análise multivariada, a estimativa do efeito do fator de uma variável altera-se com a inserção de outras variáveis no modelo, não podendo ser obtido a partir de um cálculo isolado, usando os dados de uma tabela, como a supracitada, haja vista que, a razão de chances, numa regressão logística, leva em consideração o efeito das demais variáveis previsoras incluídas no modelo (OLIVEIRA, 2009, p. 105). Ademais, a forma padrão de softwares estatísticos gerarem os coeficientes é em *logodds*, e tal valor pode ser obtido calculando o logaritmo de 0,15:

$$log\ 0{,}15 = -\ 1{,}89$$

O Varbrul/Goldvarb, no entanto, utiliza o método de codificação conhecido como *desvio da média* (cf. JOHNSON, 2009). Esse método calcula o logaritmo natural (*ln*) da *odds* (chance) de cada fator da variável previsora e, em seguida a média entre esses valores. Na sequência se verifica a diferença entre o valor de cada fator e a média e calcula-se a razão de chance ou *odds ratio* (OR) desses valores em relação à média e, aqui, reside uma grande

diferença em relação ao método empregado pelos demais softwares citados (SPSS, Rbrul e R). A razão de chances é calculada em relação à média geométrica das chances de todos os fatores da variável previsora e não em relação a um fator de referência (cf. HOSMER; LEMESHOW, 2000; OLIVEIRA, 2009). Por fim, o peso relativo é obtido a partir da *odds ratio*, da seguinte forma: *OR / 1+OR*. Além disso, as versões mais recentes do software consideram, ainda, o tamanho da interferência do fator na variável resposta a partir da quantidade de ocorrências dele (cf. MORRISON, 2005; OLIVEIRA, 2009). Já o valor do *input* é o valor do peso relativo da média das chances.

Como visto na seção 2.4.3, uma modelagem de regressão logística supõe que as observações que compõem o *corpus* sejam independentes entre si, entretanto, geralmente, não são, dado que cada informante fornece várias ocorrências (cf. JOHNSON, 2009; TAGLIAMONTE, 2012; OUSHIRO, 2017; LIMA Jr.; GARCIA, 2021). O informante é, dessa forma, uma variável que deve ser considerada na modelagem estatística, contudo, esta não é uma variável previsora fixa, replicável em outros estudos como o *Gênero* dos informantes, e sim uma variável aleatória. Ademais, os itens lexicais também se repetem nos dados e cada item lexical pode exercer diferentes efeitos sobre a variável resposta (LIMA Jr.; GARCIA, 2021).

O Goldvarb X, porém, não está equipado para analisar variáveis previsoras aleatórias. Se adicionarmos, por exemplo, a variável *Informante*, numa análise do Goldvarb, o programa irá subestimar a significância do efeito das variáveis sociais fixas – como *Gênero*, *Faixa etária*, *Nível de escolaridade e Classe social* – também referentes aos informantes, que podem ser excluídas da melhor rodada, mesmo sendo significativas. Todavia, ao não considerar a variável *Informante*, o modelo estará ignorando a influência do falante, tratando, desse modo, cada ocorrência nos dados como uma observação independente, o que superestimará o efeito das

variáveis sociais fixas (JOHNSON, 2009; TAGLIAMONTE, 2012; LIMA Jr.; GARCIA, 2021).

Para modelar variáveis fixas e aleatórias por meio um modelo de regressão logística (ou outro tipo de modelo de regressão, como a linear) misto é necessário utilizar outros softwares, como o Rbrul e o R, sobre os quais discorremos nas subseções 2.5.2 e 2.5.3.

2.5.2 O Rbrul

O Rbrul é um software gratuito, escrito por Daniel Ezra Johnson, que roda no R e na sua interface RStudio. O autor disponibiliza um manual de uso do software, facilitando, assim, sua utilização. O programa e o manual de uso estão disponíveis na página *http://www.danielezrajohnson.com/rbrul.html*. O software foi desenvolvido com a finalidade de replicar as funcionalidades do Goldvarb, calculando, inclusive, os pesos relativos, além de fornecer valores em *logodds* para os parâmetros ou coeficientes do modelo (JOHNSON, 2009, p. 362).

Destarte, o programa realiza todas as funções que o Goldvarb realiza, tais como regressão logística múltipla, tabulação cruzada e *step up* / *step down*, além de ser capaz de trabalhar com os *knockouts* sem precisar excluí-los, como o Goldvarb, e estabelece uma interface com as capacidades gráficas do R (JOHNSON, 2009; GOMES, 2012). Ademais, modela variáveis previsoras numéricas contínuas, executa modelos de regressão linear, no qual a variável resposta é numérica contínua, e modelos mistos que consideram o efeito de variáveis aleatórias como o *Informante* e o *Item lexical*.

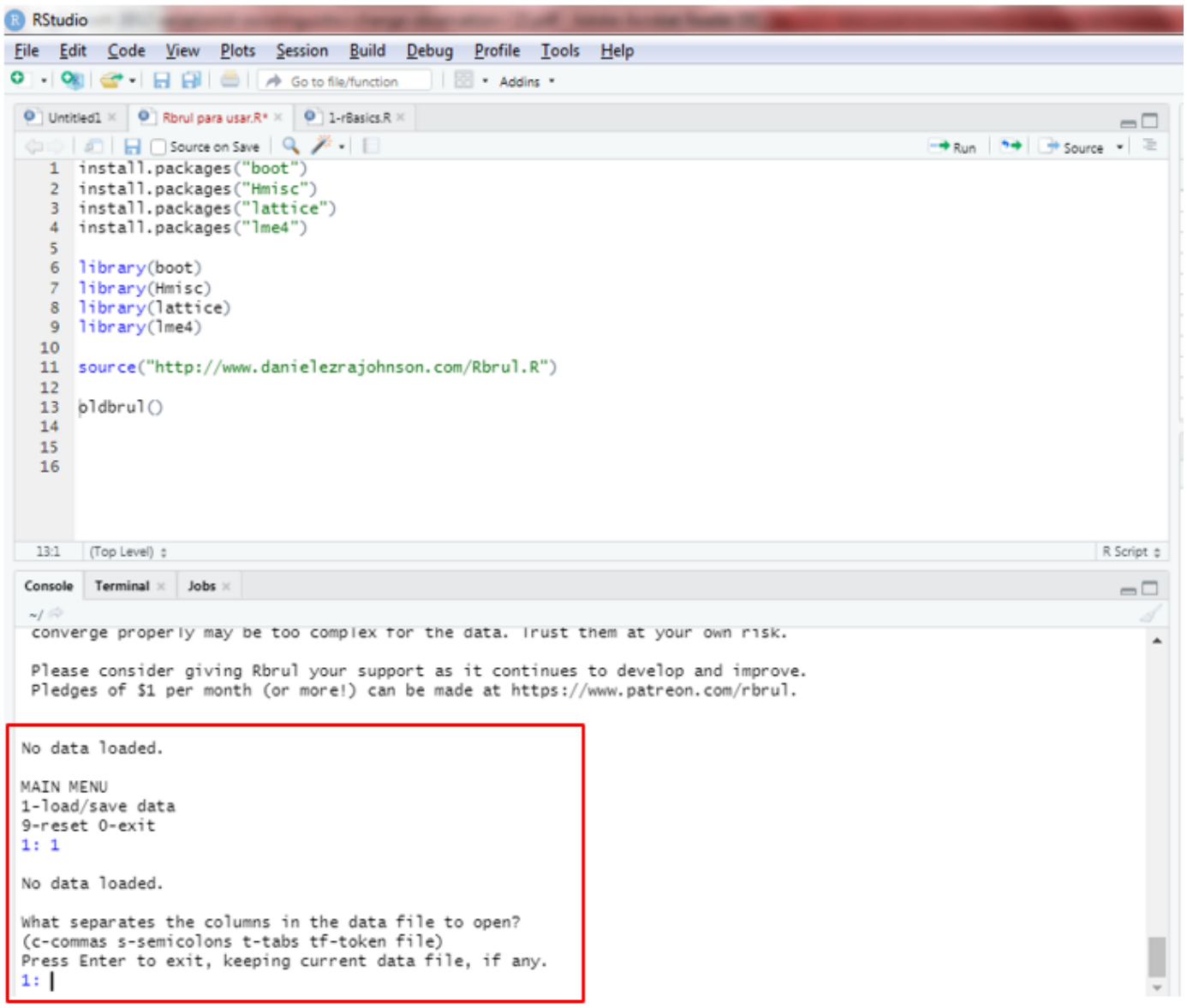
No Rbrul, um modelo misto estima valores para cada um dos níveis (ou fatores) de uma variável previsora de fixa, como *Gênero*, por exemplo. Entretanto, para variáveis aleatórias, como o *Informante*/

Falante, o modelo estima um parâmetro único representando o montante da variação entre os informantes. Diferente de um modelo de regressão ordinário que incluiria o informante como uma variável previsora fixa, um modelo misto não encaixa um parâmetro para cada informante do estudo e, justamente, devido a isso, pode captar os efeitos, de forma eficiente, de variáveis previsoras fixas como *Gênero*, *Nível de escolaridade* e *Faixa etária* dos informantes já que estes serão estimados considerando o efeito da variável *Informante* (JOHNSON, 2009, p. 362-365).

As versões anteriores eram interfaces, baseadas em texto, para as funcionalidades do software R (JOHNSON, 2009, p. 362), funcionando a partir de comandos (cf. figura 3), que guiavam o usuário por uma série de passos para executar um modelo de regressão (TAGLIAMONTE, 2012, p. 139). Para utilizar a versão atual, 3.1.4, de setembro de 2020, nesse formato, basta executar as linhas de código:

```
source("http://www.danielezrajohnson.com/Rbrul.R")
oldbrul()
```

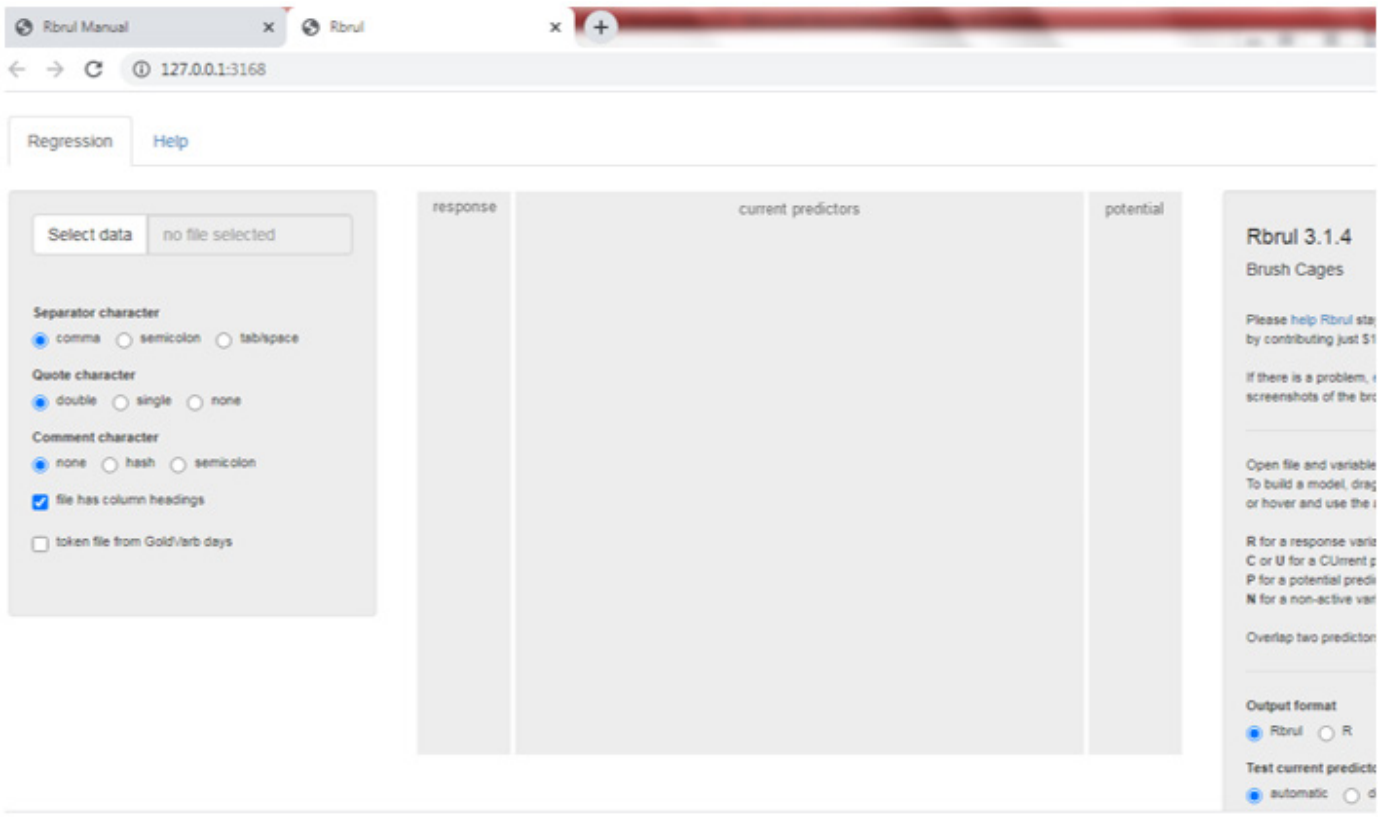
Figura 3: Rbrul: interface baseada em texto



Porém, a versão atual do programa também possui uma interface gráfica (cf. figura 4) que pode ser executada a partir das linhas:

source("http://www.danielezrajohnson.com/Rbrul.R")
rbrul()

Figura 4: Rbrul: interface gráfica

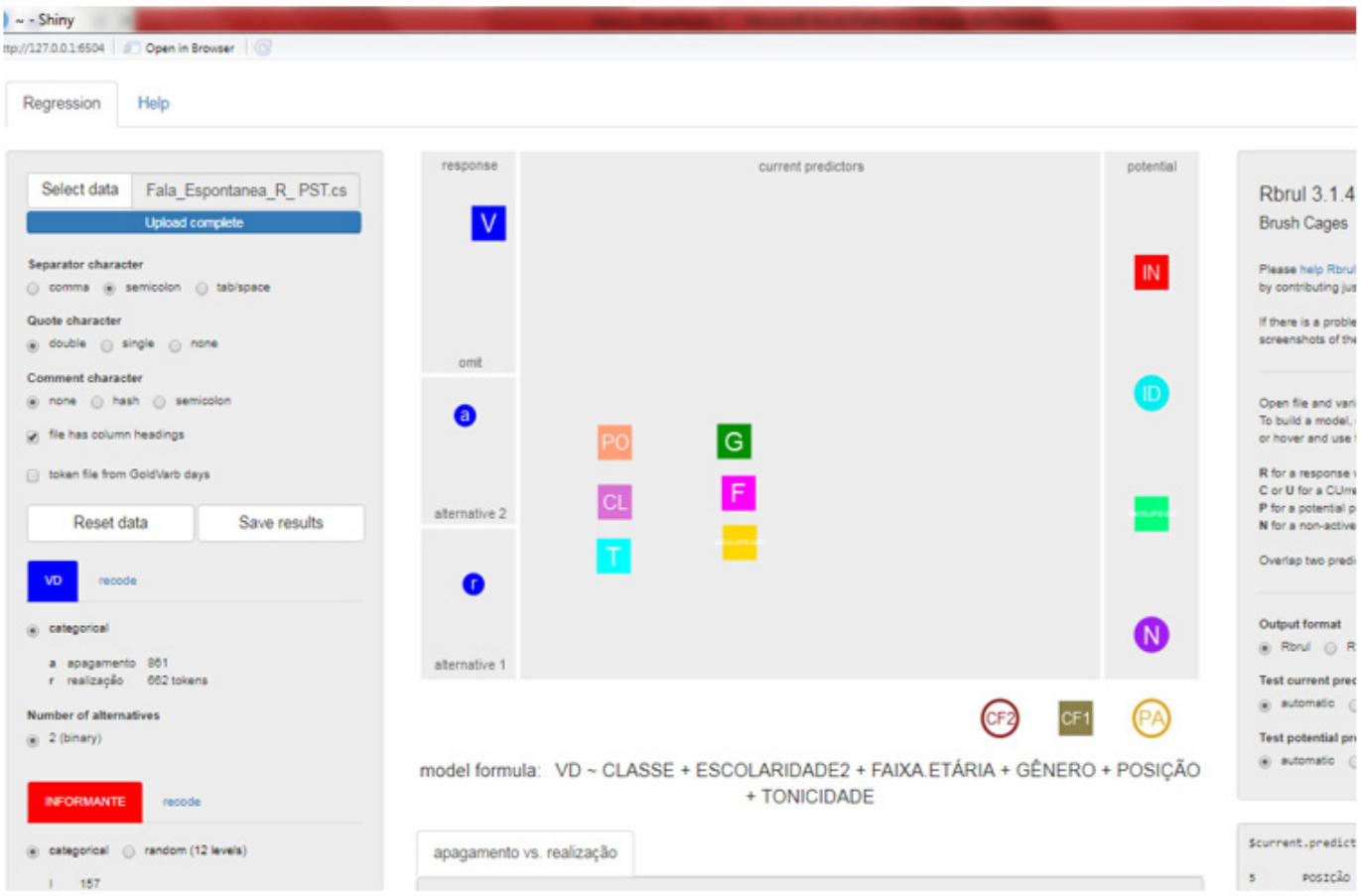


O Rbrul suporta arquivos de base de dados em vários formatos[13], no entanto, como ele não reconhece, automaticamente, o formato do arquivo, é preciso selecionar o separador de caracteres que pode ser: *comma*, *semicolon* e *tab*/*space*. O funcionamento dessa interface gráfica é bastante simples. Para executar uma regressão logística ou linear, basta selecionar a base de dados, e o separador de caracteres, sendo que se o formato do arquivo da base de dados for *.csv*, por exemplo, o separador é *semicolon*. Na sequência, as variáveis aparecerão no campo *potential*, sendo configuradas automaticamente, mas é possível reconfigurá-las na coluna esquerda, abaixo do botão *Select data*, redefinindo sua natureza (numérica ou categórica, por exemplo) e o tipo de variável (fixa ou aleatória), além de recodificar níveis de variáveis categóricas. Feito isto, é possível arrastar a variável resposta para

13 Segundo autor do código do software, o formato recomendado é *.csv*.

o campo *response* e as variáveis previsoras (fixas e aleatórias), que se quer controlar no estudo, para o campo *current predictors*, e o modelo será escrito, automaticamente. Assim, o Rbrul possibilita a inclusão de variáveis aleatórias, num modelo estatístico, sem que o pesquisador precise realizar procedimentos adicionais. Na figura 5 demonstramos a análise da regra variável de apagamento do rótico em coda no português santomense, num *corpus* obtido a partir de 12 entrevistas de fala espontânea realizadas em 2016 e 2019 em São Tomé e Príncipe (BALDUINO, 2016; 2019).

Figura 5: Rbrul: utilização da interface gráfica



Os resultados são gerados, instantaneamente, num formato de fácil interpretação bastante conhecido para o sociolinguista que estava habituado a utilizar as versões do Varbrul (cf. figura 6).

Figura 6: Rbrul: resultados

model formula: VD ~ CLASSE + ESCOLARIDADE2 + FAIXA.ETÁRIA + GÊNERO + POSIÇÃ
+ TONICIDADE

apagamento vs. realização

```
model.basics
 total.n df intercept input.prob grand.proportion
    1523  9     0.362       0.59            0.565

model.fit
 deviance      AIC     AICc Somers.Dxy   R2
 1626.243 1644.243 1644.362      0.606 0.34

CLASSE
      logodds   n proportion factor.weight
verbo    0.29 779      0.738         0.572
nome    -0.29 744      0.384         0.428

ESCOLARIDADE2
            logodds   n proportion factor.weight
fundamental   0.822 351      0.667         0.695
Médio        -0.390 654      0.616         0.404
Superior     -0.432 518      0.432         0.394

FAIXA.ETÁRIA
         logodds   n proportion factor.weight
primeira   0.741 531      0.661         0.677
segunda   -0.304 618      0.542         0.425
Terceira  -0.437 374      0.468         0.392

GÊNERO
          logodds   n proportion factor.weight
Feminino    0.204 821      0.622         0.551
Masculino  -0.204 702      0.499         0.449

POSIÇÃO
          logodds   n proportion factor.weight
final       0.895 811      0.776          0.71
não-final  -0.895 712      0.326          0.29

TONICIDADE
       logodds    n proportion factor.weight
tônica  0.0368 1066      0.660         0.509
átona  -0.0368  457      0.344         0.491
```

Inicialmente o *output* apresenta o total de ocorrências analisadas (*total.n*=1523), um valor de *intercept* (0,362) para o modelo de regressão, o *input* em probabilidade (.59), como o Goldvarb, e a proporção de aplicação da regra variável (0,565, equivalente a 56,5%), nesse caso, o apagamento do rótico. O valor do *intercept* é o valor de referência para o cálculo dos coeficientes

em *logodds* e o *input* é uma probabilidade geral de aplicação da regra. Na sequência, são apresentadas algumas medidas estatísticas, sobre o modelo, como o *AIC* (*Akaike Information Criterion*), uma medida que permite, além da mensuração da qualidade do modelo estatístico, a comparação entre diferentes modelos.

Quanto ao valor do efeito dos fatores das variáveis previsoras, o Rbrul apresenta, como o Goldvarb, os valores dos coeficientes em probabilidade, denominados pela Sociolinguística, pesos relativos (*factor.weight*), para cada um dos fatores ou níveis das variáveis, mas também apresenta esses valores em *logodds*. Além disso, calcula o número total de ocorrências (*n*) e a proporção de aplicação da regra (*proportion*) para cada fator analisado.

Na coluna à esquerda o Rbrul apresenta as variáveis previsoras incluídas no modelo (*$current.predictors*), reordenadas pelo nível de significância – da mais significante para a menos significante. O nível de significância é indicado pelo valor-p (*p.value*), tratado, de forma mais detida, na seção 2.4.2. Nesse *output*, quanto menor o valor-p maior é o nível de significância da variável. Convencionalmente, a comunidade científica utiliza o limite máximo de 0,05, para considerar algo como pouquíssimo provável de acontecer ao acaso, dessa forma, um valor-p acima de 0,05 indica que a hipótese nula – segundo qual a variável não tem um efeito estatisticamente significativo sobre a aplicação da regra variável – não pode ser rejeitada. Com exceção da variável *Tonicidade da sílaba*, as demais variáveis previsoras têm um efeito significativo sobre a aplicação da regra de apagamento do rótico.

Por fim, o Rbrul calcula, automaticamente, o nível de significância das demais variáveis previsoras (*$potential.predictors*) contidas na base dados – não incluídas no modelo – e das interações entre as variáveis previsoras incluídas no modelo, organizando-as, também, de acordo com sua significância, conforme a figura 7.

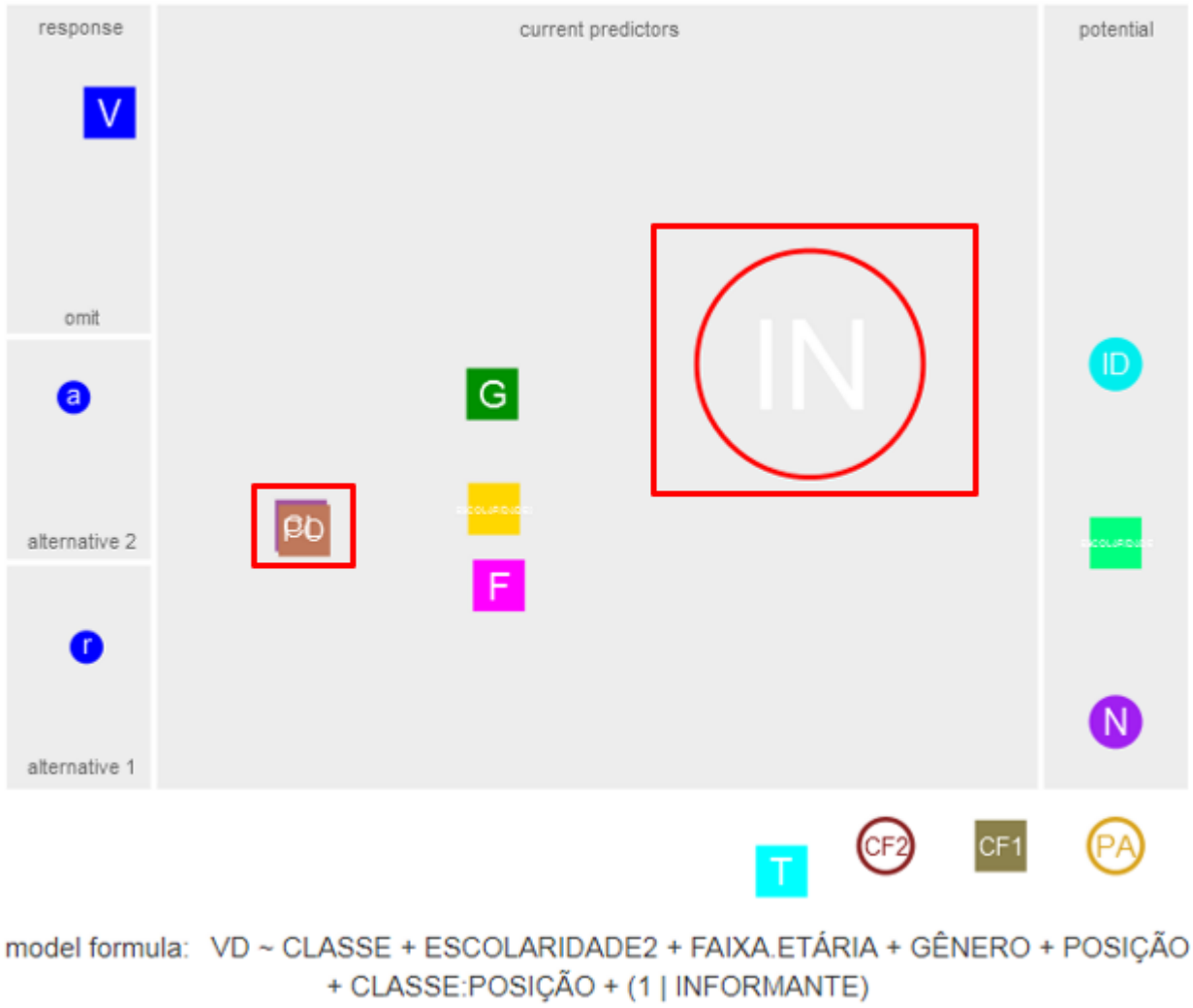
Figura 7: Rbrul: resultados – significância das variáveis

```
$current.predictors
                df AIC.if.dropped  p.value
5       POSIÇÃO  1         +95.71 1.50e-26
2 ESCOLARIDADE2  2         +43.23 5.05e-14
1        CLASSE  1          -1.81 5.72e-05
3  FAIXA.ETÁRIA  2          +0.51 9.57e-05
4        GÊNERO  1          -9.02 2.73e-03
6    TONICIDADE  1         -17.79     0.65
```

```
$potential.predictors
                             df AIC.if.added  p.value
3                 INFORMANTE  7      -108.59 9.37e-17
1               ESCOLARIDADE  3       -85.68 1.34e-14
10 ESCOLARIDADE2:FAIXA.ETÁRIA  1       -43.66 4.06e-07
11       ESCOLARIDADE2:GÊNERO  2       -45.24 1.21e-06
8              CLASSE:POSIÇÃO  1       -26.73 3.13e-03
15       FAIXA.ETÁRIA:POSIÇÃO  2       -28.97 4.16e-03
9           CLASSE:TONICIDADE  1       -25.19 7.34e-03
12      ESCOLARIDADE2:POSIÇÃO  2       -24.61   0.0367
6         CLASSE:FAIXA.ETÁRIA  2       -21.37    0.185
19         POSIÇÃO:TONICIDADE  1       -19.52    0.218
13   ESCOLARIDADE2:TONICIDADE  2       -20.73    0.255
16    FAIXA.ETÁRIA:TONICIDADE  2       -20.53    0.282
5        CLASSE:ESCOLARIDADE2  2       -20.19    0.335
17             GÊNERO:POSIÇÃO  1        -18.5     0.48
2                      IDADE  1       -18.45    0.502
4                  Nº.ARQUIVO  1       -18.45    0.502
18          GÊNERO:TONICIDADE  1       -18.14    0.705
14       FAIXA.ETÁRIA:GÊNERO  2       -18.39    0.822
7                CLASSE:GÊNERO  1       -18.05     0.83
```

Para incluir uma interação no modelo basta sobrepor duas variáveis previsoras no campo *current predictors*. Aqui, incluímos, no modelo, a interação entre as variáveis CLASSE da palavra e POSIÇÃO do segmento na palavra (contornada em vermelho), além de uma variável aleatória INFORMANTE (contornada em vermelho), empregando um modelo misto, conforme a figura 8.

Figura 8: Rbrul: utilização da interface gráfica – incluindo uma interação e uma variável aleatória (modelo misto)



Com a utilização de um modelo misto, incluindo a variável aleatória INFORMANTE, a significância das variáveis fixas é recalculada (cf. figura 9) e as variáveis sociais GÊNERO e FAIXA. ETÁRIA que no modelo anterior – sem a inclusão da variável aleatória em questão – eram relevantes para a aplicação da regra variável, nesse modelo, deixam de sê-lo, indicando que ao não incluir a variável aleatória INFORMANTE, a significância do efeito das variáveis sociais estava superestimada, isto é, o efeito observado, anteriormente, se deve a alguns informantes e considerando-os, como variável aleatória, chegamos à conclusão de que as variáveis GÊNERO e FAIXA.ETÁRIA não têm um efeito verdadeiro sobre a aplicação da regra.

Figura 9: Rbrul: resultados – significância das variáveis (modelo misto com interação)

```
$current.predictors
                 df AIC.if.dropped  p.value
4 CLASSE:POSIÇÃO  1          +7.95 1.61e-03
1  ESCOLARIDADE2  2          +2.65   0.0359
2   FAIXA.ETÁRIA  2          +0.03    0.134
3         GÊNERO  1          -0.53    0.225
```

```
$potential.predictors
                            df AIC.if.added  p.value
1                ESCOLARIDADE  3       -11.64 5.21e-04
7  ESCOLARIDADE2:FAIXA.ETÁRIA  1         -1.7   0.0544
11       FAIXA.ETÁRIA:POSIÇÃO  2         -0.9   0.0864
8        ESCOLARIDADE2:GÊNERO  2        -0.11    0.128
9       ESCOLARIDADE2:POSIÇÃO  2        +0.19    0.149
5         CLASSE:FAIXA.ETÁRIA  2        +0.74    0.196
6               CLASSE:GÊNERO  1        +1.06    0.333
3                  Nº.ARQUIVO  1        +1.73    0.602
12            GÊNERO:POSIÇÃO  1        +1.87    0.716
4       CLASSE:ESCOLARIDADE2  2        +3.59    0.813
2                      IDADE  1        +1.98    0.876
10       FAIXA.ETÁRIA:GÊNERO  2        +3.95    0.975
```

Os resultados, na figura 10, podem ser interpretados da mesma forma que na figura 6, com o acréscimo da interação e do resultado para a variável aleatória INFORMANTE. A interação mostra os totais de ocorrências, os percentuais e os pesos relativos (*factor.weight*), do cruzamento das variáveis POSIÇÃO e CLASSE. Já para a variável aleatória INFORMANTE, justamente por ser uma variável aleatória, em vez de calcular um valor de peso relativo para cada um dos informantes, o software calcula apenas um valor de referência (*intercept*) para a variável, contudo, assim como ocorre com significância, o efeito das variáveis fixas (peso relativo) é recalculado considerando o efeito da variável INFORMANTE.

Figura 10: Rbrul: resultados (modelo misto com interação)

apagamento vs. realização

```
model.basics
 total.n df intercept input.prob grand.proportion
    1523 10     0.349      0.586            0.565

model.fit
 deviance      AIC     AICc Somers.Dxy.fixed Somers.Dxy.total R2.fixed R2.total
 1563.231 1583.231 1583.376            0.612            0.656    0.363    0.421

CLASSE
      logodds   n proportion factor.weight
verbo   0.295 779      0.738         0.573
nome   -0.295 744      0.384         0.427

ESCOLARIDADE2
            logodds   n proportion factor.weight
fundamental   0.908 351      0.667         0.713
Superior     -0.398 518      0.432         0.402
Médio        -0.510 654      0.616         0.375

FAIXA.ETÁRIA
         logodds   n proportion factor.weight
primeira   0.958 531      0.661         0.723
segunda   -0.414 618      0.542         0.398
Terceira  -0.544 374      0.468         0.367

GÊNERO
          logodds   n proportion factor.weight
Feminino    0.248 821      0.622         0.562
Masculino  -0.248 702      0.499         0.438

POSIÇÃO
          logodds   n proportion factor.weight
final       0.988 811      0.776         0.729
não-final  -0.988 712      0.326         0.271

CLASSE:POSIÇÃO interaction
                logodds   n proportion factor.weight
nome:não-final    0.233 573      0.319         0.558
verbo:final       0.233 640      0.822         0.558
nome:final       -0.233 171      0.602         0.442
verbo:não-final  -0.233 139      0.353         0.442

INFORMANTE
        intercept    n proportion
std.dev     0.573 1523      0.565
```

Essa ferramenta estabelece uma interface com algumas das funcionalidades do R, no entanto, não permite a realização de tarefas de manipulação da base de dados, como, por exemplo, a criação de novas variáveis e a filtragem de dados, entre outras. Para ter acesso a todas as funcionalidades do R é preciso utilizar seu próprio ambiente (RStudio), que apresentaremos na subseção 2.5.3.

2.5.3 O R e sua interface RStudio

O R é um software especializado em manipulação, análise e visualização gráfica de dados estatísticos que utiliza uma linguagem de programação homônima.[14] Além disso, é um software expansível graças à possibilidade de utilização dos chamados *packages* (cf. GRIES, 2013) com dados e funções para diferentes áreas do conhecimento relacionado à estatística sendo, atualmente, considerado um dos melhores ambientes computacionais para o tratamento de dados estatísticos. Uma de suas vantagens é o fato de o software ser gratuito e estar disponível para uma variedade de plataformas (Unix, Windows e MacOS) sob os termos da Licença Pública Geral GNU da *Free Software Foundation* (cf. FERREIRA, 2013).

O R começou a ser desenvolvido por Robert Gentleman e Ross Ihaka ("R & R"), ambos do Departamento de Estatística da University of Auckland, Nova Zelândia, em 1991, e o primeiro relato de distribuição foi em 1993, quando algumas cópias foram disponibilizadas no *StatLib*, um sistema de distribuição de softwares estatísticos. Em 1995 Martin Mächler (do Instituto Federal de Tecnologia de Zurique, na Suíça), "R & R", lançou o código fonte do R, e em 1997 foi criado um grupo de profissionais com a tarefa de atualizar o código, possibilitando, assim, a atualização mais rápida do software. Desde então, o R vem sendo cada vez mais utilizado em todo o mundo (cf. MELO, 2017).

O RStudio é uma interface funcional e mais amigável para o R, sendo o principal ambiente de desenvolvimento integrado para R, que disponibiliza ferramentas adicionais diretamente na

14 A linguagem de programação R é voltada para a análise de dados, e pode ser utilizada para realizar computações estatísticas e gráficas, compilar corpora, produzir listas de frequências, entre outras diversas tarefas (OUSHIRO, 2014:134-136).

interface gráfica, tais como a visualização dos *scripts*[15] abertos recentemente, o histórico de linhas de comando executadas, a lista de pacotes instalados, entre outras (OUSHIRO, 2014:136).

Além de executar vários modelos estatísticos, como a regressão linear, e os demais modelos da família de modelos lineares generalizados (Regressão Logística, Ordinal, Poisson, Multinomial, etc.), o R analisa, também, variáveis previsoras de efeitos aleatórios por meio de modelos mistos, executando, ainda, diversas outras tarefas, como elaboração de diversos gráficos, tabelas e cálculos de medidas estatísticas, tais como, média, mediana, desvio padrão, variância etc. (cf. GRIES, 2013; LEVSHINA, 2015; OUSHIRO, 2017; WINTER, 2020).

Na interface do R, RStudio, os comandos são efetuados por meio de linhas de códigos em linguagem R que constituem *scripts*. Os *scripts* podem ser usados para importar, manipular, visualizar e analisar dados. O primeiro passo do *script* é definir o diretório em que está localizada a base de dados, em seguida, carregar a base de dados, que é o mesmo que solicitar ao R que leia a base de dados – que pode ser um arquivo em formato em formato *.txt*, *.csv*, *.xls* ou *.xlsx*, entre outros formatos – criando um objeto no *Environment* do RStudio. É importante verificar a estrutura dos dados do novo objeto criado por meio de uma inspeção. Se houver necessidade de manipular a base de dados (como renomear variáveis ou níveis de uma variável, criar novas variáveis, filtrar dados etc.) é possível fazer isso no R, de forma rápida e simples, por meio, da linguagem básica do R, e da linguagem do pacote *Tidyverse*. Vejamos as linhas de códigos, de um *script*, utilizadas para realizar algumas das tarefas citadas, analisando a regra variável de apagamento do rótico em coda no português santomense utilizando uma base de dados constituída a partir de 12 entrevistas de fala espontânea realizadas

15 Em informática, um *script* é um conjunto de instruções para que uma função seja executada em determinado software.

em 2016 e 2019 em São Tomé e Príncipe (BALDUINO, 2016; 2019):

```
#definição do diretório de trabalho[16]
setwd("C:/Users/Nancy/Dropbox/R")

#carregando a base de dados e criando o objeto: "dados"
dados <- read.csv("Fala_Espontanea_R..csv", header = T, sep = ",")

#inspecionando dados
str(dados)

#codificação de dados

#convertendo as variáveis do tipo "character" (padrão do RStudio) em "factor"
dados$VD <- as.factor(dados$VD)
dados$GÊNERO <- as.factor(dados$GÊNERO)
dados$ESCOLARIDADE <- as.factor(dados$ESCOLARIDADE)
dados$CF1 <- as.factor(dados$CF1)
dados$CF2 <- as.factor(dados$CF2)
dados$TONICIDADE <- as.factor(dados$TONICIDADE)
dados$CLASSE <- as.factor(dados$CLASSE)
dados$POSIÇÃO <- as.factor(dados$POSIÇÃO)
dados$PALAVRA <- as.factor(dados$PALAVRA)
dados$INFORMANTE <- as.factor(dados$INFORMANTE)

#renomeando níveis de uma variável
levels(dados$ESCOLARIDADE)<-list("Ensino  fundamental" =
```

16 O conteúdo escrito após o símbolo # não faz parte das linhas de códigos do R, é meramente explicativo. Esse símbolo é uma forma de avisar o R que não estamos digitando nenhum comando, mas apenas alguma nota explicativa.

```
"4ª classe",
                                "Ensino fundamental" = "9ª classe",
                                "Ensino médio" = "10ª classe",
                                "Ensino médio" = "12ª classe",
                                "Superior" = "Ensino Superior",
                                "Superior" = "Mestrado")

    # carregando o pacote Tidyverse
    library(tidyverse)

    #criando uma nova variável (FAIXA.ETÁRIA) na base de dados
com 3 níveis: "1a", "2a" e "3a"
    dados1 <- dados1 %>%
     mutate(FAIXA.ETÁRIA = if_else (IDADE <= 20, "1a", if_else
     (IDADE>=21 & IDADE<=40, "2a", "3a")))

    #inspecionando dados
    str(dados)
```

O código *str()* exibe a estrutura da base de dados que inclui informações como o número de observações e as variáveis com seus respectivos níveis conforme a figura 11.

Figura 11: Estrutura dos dados no RStudio

```
'data.frame':   1523 obs. of  15 variables:
 $ X           : logi  NA NA NA NA NA NA ...
 $ VARIEDADE   : Factor w/ 1 level "PST": 1 1 1 1 1 1 1 1 1 1 ...
 $ VD          : Factor w/ 2 levels "apagamento","realização": 1 1 1 1 1 1 2 2 2 1 .
 $ INFORMANTE  : Factor w/ 12 levels "I","II","III",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ GÊNERO      : Factor w/ 2 levels "Feminino","Masculino": 1 1 1 1 1 1 1 1 1 1 ...
 $ IDADE       : int  33 33 33 33 33 33 33 33 33 33 ...
 $ FAIXA.ETÁRIA: Factor w/ 3 levels "1a","2a","3a": 2 2 2 2 2 2 2 2 2 2 ...
 $ ESCOLARIDADE: Factor w/ 3 levels "Ensino Fundamental",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Nº.ARQUIVO  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ PALAVRA     : Factor w/ 534 levels "(es)tiver","abaixar",..: 262 226 22 116 292 2
 $ CF1         : Factor w/ 8 levels "#","a","e","e ab",..: 3 3 2 5 2 7 6 8 4 3 ...
 $ CF2         : Factor w/ 14 levels "#","b","d","g",..: 1 1 1 1 1 1 1 8 1 1 ...
 $ TONICIDADE  : Factor w/ 2 levels "átona","tônica": 2 2 2 2 2 2 2 2 2 2 ...
 $ CLASSE      : Factor w/ 2 levels "nome","verbo": 2 2 2 2 2 1 1 1 2 2 ...
 $ POSIÇÃO     : Factor w/ 2 levels "final","não-final": 1 1 1 1 1 1 1 2 1 1 ...
```

Vejamos as linhas de códigos que realizam uma regressão logística considerando as variáveis *Posição do segmento*, *Classe gramatical da palavra*, *Tonicidade da sílaba*, *Gênero*, *Faixa etária*} e *Escolaridade* dos informantes:

#modelo de regressão logística
regres_log<-glm(VD~ POSIÇÃO + CLASSE + TONICIDADE + GÊNERO + FAIXA.ETÁRIA + ESCOLARIDADE, data = dados, family = binomial)

Para incluir interações no modelo, basta substituir o sinal de adição "+" por um asterisco "*" :

#modelo de regressão logística com interação

*regres_log<-glm(VD~ POSIÇÃO * CLASSE + TONICIDADE + GÊNERO + FAIXA.ETÁRIA + ESCOLARIDADE, data = dados, family = binomial)*
summary(regres_log) #exibe um resumo dos resultados da regressão

Um dos maiores desafios de se utilizar o RStudio é interpretar os resultados estatísticos gerados, destarte, apresentamos, na figura 12, o *summary*, ou resumo, do resultado da regressão e, na sequência, explicitamos os resultados fornecidos.

Figura 12: *Summary* da regressão logística com interação no RStudio

```
Call:
glm(formula = VD ~ POSIÇÃO * CLASSE + TONICIDADE + GÊNERO + FAIXA.ETÁRIA +
    ESCOLARIDADE, family = binomial, data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9750  -0.8475  -0.4484   0.9317   2.2339

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -2.426312   0.373475  -6.497 8.22e-11 ***
POSIÇÃOnão-final                1.440787   0.207564   6.941 3.88e-12 ***
CLASSEverbo                    -0.986231   0.196877  -5.009 5.46e-07 ***
TONICIDADEtônica               -0.008031   0.162256  -0.049 0.960522
GÊNEROMasculino                 0.421354   0.137120   3.073 0.002120 **
FAIXA.ETÁRIA2a                  1.011432   0.246643   4.101 4.12e-05 ***
FAIXA.ETÁRIA3a                  1.110011   0.314614   3.528 0.000418 ***
ESCOLARIDADEEnsino Médio        1.173861   0.278068   4.221 2.43e-05 ***
ESCOLARIDADEEnsino Superior     1.251010   0.170021   7.358 1.87e-13 ***
POSIÇÃOnão-final:CLASSEverbo    0.850473   0.288420   2.949 0.003191 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2085.2  on 1522  degrees of freedom
Residual deviance: 1617.5  on 1513  degrees of freedom
AIC: 1637.5

Number of Fisher Scoring iterations: 4
```

Inicialmente, o RStudio informa a linha de código com a fórmula empregada, nesse caso a função *glm* que executa uma regressão logística. Em seguida, temos a distribuição dos resíduos em quartis.[17] Esses resíduos são a diferença entre o valor previsto pelo modelo e o valor observado. O Ideal é que o valor da mediana esteja próximo de zero e os valores *min-max* e *1Q-3Q* são razoavelmente simétricos. Isso porque os resíduos são os valores que o modelo não foi capaz de prever perfeitamente e é normal que haja resíduos num modelo, mas se um modelo `erra' é bom que ele `erre' tanto para mais quanto para menos, a fim de que possamos "ter certa segurança de que nossas estimativas não estão muito longe do que

17 Quartis são valores que dividem uma amostra de dados em quatro partes iguais. Com eles é possível avaliar a dispersão e a tendência central de um conjunto de dados. No 1º quartil (1Q) 25% dos dados são menores ou iguais a esse valor. O 2º quartil é a *mediana* e 50% dos dados são maiores que esse valor e 50% são menores. No 3º quartil (3Q) 75% dos dados são menores que esse valor (OUSHIRO, 2017).

se poderia observar" (OUSHIRO, 2017, p. 135).

Depois o RStudio mostra os parâmetros, ou coeficientes, de cada nível ou fator de uma variável previsora e do *intercept*. O valor do *intercept* é o coeficiente linear e, também, o valor de referência para o cálculo dos coeficientes dos fatores de cada variável. Os valores nas linhas seguintes são os coeficientes angulares, as estimativas do valor do efeito de cada um dos fatores sobre a aplicação da regra variável, inclusive dos fatores das variáveis em interação, a qual, de acordo com o valor-p, é significativa. Os valores são calculados em *logodds*, todavia, é possível convertê-los em valores de probabilidade (cf. GRIES, 2013; OUSHIRO, 2017). No RStudio, essa conversão pode ser feita utilizando uma função que não faz parte do R, mas pode ser incorporada com as linhas de comandos:

```
ilogit <- function(x) {
  1/(1+exp(-x))
  ilogit(1.440787)
```

É importante notar que o primeiro nível de cada variável não aparece na lista de coeficientes estimados, o que acontece porque este é o fator de referência, a partir do qual, os valores dos coeficientes dos demais fatores da variável serão calculados. Os valores dos coeficientes dos fatores de referência estão contidos no valor do *intercept*.

A regressão logística apresenta o erro padrão junto à estimativa de cada coeficiente. Quando se faz a divisão entre Estimativa/Erro Padrão, chega-se ao valor da terceira coluna, o *valor-z*, que pode ser consultado numa tabela de distribuição normal padrão a partir da qual se pode obter o valor do nível de significância (OUSHIRO, 2017, p. 188). No entanto esse valor já é fornecido pelo R na próxima coluna e corresponde ao valor-p sobre o qual discorremos, com maior nível de detalhamento, na seção 2.4.2. Diferente do Goldvarb e do

Rbrul que fornecem um valor-p para indicar o nível de significância de cada variável previsora, o R, fornece um valor-p para cada nível, ou fator e, para que o valor do efeito de cada nível de uma variável seja considerado estatisticamente significativo, o valor-p deve ser menor do que o nível α preestabelecido. Convencionalmente, esse nível é 0,05, mas o pesquisador pode adotar outro valor. Em suma, quanto menor for o valor-p, maior é o nível de significância do efeito do fator que está sendo testado. De acordo com os resultados, na figura 12, à exceção dos fatores da variável *Tonicidade*, o efeito dos níveis de todas as demais variáveis é significativo para a aplicação da regra.

Abaixo da tabela de coeficientes, o RStudio mostra o significado dos asteriscos, de acordo com os níveis de significância mais comuns: 0.001, 0.01 e 0.05. Três asteriscos indicam que $p <$ 0.001, dois asteriscos indicam que *p* está entre 0.001 e 0.01, e um asterisco indica que *p* está entre 0.01 e 0.05. O *ponto final* indica um valor um pouco acima de 0.05 (OUSHIRO, 2017).

O desvio nulo se refere à variabilidade total dos dados, antes da inclusão de qualquer variável previsora e o desvio residual se refere à variabilidade nos dados depois da inclusão das variáveis previsoras. Portanto, a diferença entre o desvio residual e o desvio nulo é o quanto o nosso modelo é capaz de prever da variabilidade dos dados. O AIC (*Akaike Information Criterion*) mede a qualidade do modelo podendo ser utilizado para comparar modelos. A função *step( )*, inclusive, se baseia no valor dessa medida para selecionar ou excluir variáveis. O *Fisher Scoring iterations* se refere ao número de iterações do modelo até que os resultados tenham convergido. Em nosso exemplo, o modelo convergiu após 4 tentativas. Quando este número é superior a 20, é um indicativo de que foram incluídas mais variáveis previsoras do que é possível explicar com a quantidade de dados de que se dispõe (OUSHIRO, 2017, p. 189).

Para utilizar um modelo misto, que inclui variáveis aleatórias,

como o *Informante* e o *Item lexical*, utiliza-se a função *glmer* e, para tanto, é necessário carregar dois pacotes: *lmer* e *lmer*Test, executando as seguintes linhas de códigos:

library(lme4) #carrega o pacote "lmer"
library(lmerTest) #carrega o pacote "lmerTest"

*EFEITOS.MISTOS_log<-glmer (VD ~ POSIÇÃO * CLASSE + TONICIDADE + GÊNERO + FAIXA.ETÁRIA + ESCOLARIDADE + (1|INFORMANTE) + (1|PALAVRA), data = dados, family = binomial)*
summary(EFEITOS.MISTOS_log)

Figura 13: Summary do modelo de regressão logística mista no RStudio

```
> summary(EFEITOS.MISTOS_log)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: VD ~ POSIÇÃO * CLASSE + TONICIDADE + GÊNERO + FAIXA.ETÁRIA +
    ESCOLARIDADE + (1 | INFORMANTE) + (1 | PALAVRA)
   Data: dados

     AIC      BIC   logLik deviance df.resid
  1559.4   1623.4   -767.7   1535.4     1511

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3527 -0.5391 -0.1798  0.5126  6.1807

Random effects:
 Groups     Name        Variance Std.Dev.
 PALAVRA    (Intercept) 0.7160   0.8462
 INFORMANTE (Intercept) 0.4179   0.6465
Number of obs: 1523, groups:  PALAVRA, 534; INFORMANTE, 12

Fixed effects:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -3.14958    0.93186  -3.380 0.000725 ***
POSIÇÃOnão-final               1.62991    0.30341   5.372 7.79e-08 ***
CLASSEverbo                   -1.24059    0.27641  -4.488 7.18e-06 ***
TONICIDADEtônica              -0.06085    0.23319  -0.261 0.794133
GÊNEROMasculino                0.55743    0.44693   1.247 0.212307
FAIXA.ETÁRIA2a                 1.52685    0.75346   2.026 0.042720 *
FAIXA.ETÁRIA3a                 1.68900    0.97950   1.724 0.084646 .
ESCOLARIDADEEnsino Médio       1.53584    0.85952   1.787 0.073960 .
ESCOLARIDADEEnsino Superior    1.47758    0.54444   2.714 0.006649 **
POSIÇÃOnão-final:CLASSEverbo   1.04738    0.38962   2.688 0.007184 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) POSIÇÃOn- CLASSE TONICI GÊNERO FAIXAETARIA2 FAIXAETARIA3 ESCOLM ESCOLS
POSIÇÃOn-fn -0.357
CLASSEverbo -0.187  0.609
TONICIDADEt -0.237  0.365    -0.072
GÊNEROMscln  0.104  0.008    -0.002 -0.019
```

As informações sobre os resíduos e as variáveis fixas, na figura

13, podem ser interpretadas da mesma forma que na figura 12, mas nesse modelo temos, também, os resultados para as variáveis aleatórias (*Random effects*) às quais, como no Rbrul, é atribuído apenas um valor de referência (*intercept*) que altera a estimativa do efeito dos fatores das varáveis fixas e sua significância. Na verdade, segundo Oushiro (2017, p. 177), o principal resultado a checar num modelo misto é se os mesmos fatores das variáveis fixas continuam a ser relevantes após a inclusão das variáveis aleatórias. Nesse caso, o *gênero masculino*, o *nível de escolaridade ensino médio* e a *3ª faixa etária*, significativos no modelo anterior, deixaram de sê-lo indicando que tais variáveis não têm um efeito verdadeiro sobre a variável resposta.

2.6 Síntese do capítulo

Neste capítulo apresentamos a Sociolinguística Variacionista, uma ciência interdisciplinar que tem como objeto de estudo a língua, como é usada na vida cotidiana de uma comunidade, considerando os fatores sociais que se correlacionam a ela (cf. LABOV, 1972). A seguir, na seção 2.2, discorremos os pressupostos teóricos da Teoria da Variação e Mudança, conforme Weinreich, Labov e Herzog (2006[1968]), que estuda a língua como um objeto, constituído de heterogeneidade ordenada, que muda de acordo com as mudanças ocorridas na estrutura social de uma comunidade de fala. Segundo os autores, explicar a mudança linguística depende da possibilidade de descrever a diferenciação ordenada dentro da língua, uma vez que toda mudança implica variabilidade e heterogeneidade.

Na seção 2.3, apresentamos as etapas da metodologia da Sociolinguística Variacionista: (i) identificação da variável linguística a ser analisada; (ii) os critérios de seleção dos informantes e a escolha da comunidade da fala; (iii) o trabalho de campo, durante a coleta de dados; e (iv) o tratamento quantitativo dos dados,

mostrando a evolução dos modelos matemáticos empregados pela Sociolinguística Variacionista desde a década de 1960 até os dias atuais. O modelo utilizado atualmente, e o mais recomendado para análises estatísticas em estudos sociolinguísticos, é o de regressão logística mista, que é capaz de modelar variáveis previsoras de efeitos fixos e aleatórios, como *Informante* e *Item lexical*.

Na seção 2.4, introduzimos algumas noções básicas de estatística sobre probabilidade, chance (*odds*) e razão de chances (*odds ratio*), além de testes de significância utilizados para testar hipóteses, quando explicitamos os conceitos de hipótese nula (H_0) e hipótese alternativa (H_1). Esta última é a hipótese que está sendo testada, como, por exemplo, a afirmação de que há uma relação entre duas variáveis, enquanto a H_0, normalmente, é formulada como a negação da H_1, afirmando que não há relação entre as variáveis e que a distribuição dos dados observada resulta de uma flutuação aleatória e/ou erro de amostragem. Também discorremos sobre modelos estatísticos, nomeadamente o de regressão linear e o de regressão logística, o primeiro utilizado para analisar variáveis respostas numéricas (contínuas), com base em dados que seguem uma distribuição normal, e o segundo para analisar variáveis respostas binárias utilizando dados que seguem uma distribuição homônima.

Na seção 2.5, tratamos da utilização dos softwares Varbrul/Goldvarb, Rbrul e R, por meio de sua interface RStudio, que podem ser utilizados para realizar os cálculos dos parâmetros de um modelo de regressão logística (incluindo, ou não, variáveis aleatórias), apontando as limitações e/ou vantagens de utilizar cada uma dessas ferramentas.

O software Varbrul é um pacote de programas que foi criado especialmente para conduzir análises estatísticas em estudos sociolinguísticos, especificamente, para realizar análises de regras variáveis que controlam variáveis linguísticas binárias. A ferramenta

não exige muito conhecimento na área de estatística e fornece resultados, num formato, com o qual, os sociolinguistas estão habituados a trabalhar, contudo, não está equipado para analisar variáveis previsoras aleatórias, haja vista que não executa modelos mistos.

Dessa forma, se adicionarmos, por exemplo, a variável *Informante*, numa análise do Goldvarb, o programa irá subestimar a significância do efeito de variáveis sociais fixas como *Classe social*, *Gênero*, *Faixa etária* e *Nível de escolaridade*, também referentes aos informantes, que podem ser excluídas da melhor rodada, mesmo sendo significativas. Todavia, ao não considerar a variável *Informante*, o modelo estará ignorando a influência do falante, tratando, desse modo, cada ocorrência nos dados como uma observação independente, o que superestimará o efeito das variáveis sociais fixas (JOHNSON, 2009, p. 363-364; TAGLIAMONTE, 2012). Para modelar variáveis previsoras fixas e aleatórias é necessário utilizar um software que execute modelos de regressão mistos como o Rbrul e o R.

O Rbrul foi desenvolvido com a finalidade de replicar as funcionalidades do Goldvarb, calculando, inclusive, os pesos relativos, além de fornecer valores em *logodds* (JOHNSON, 2009). Ademais, modela variáveis previsoras numéricas contínuas, executa modelos de regressão linear, nos quais a variável resposta é numérica (contínua) e modelos mistos, que incluem variáveis aleatórias. O software roda no R e estabelece uma interface com algumas das suas funcionalidades, no entanto, não permite a realização de tarefas de manipulação da base de dados, como, por exemplo, a criação de novas variáveis, filtragem de dados, entre outras. Para ter acesso a todas as funcionalidades do R é preciso utilizar sua própria interface, a saber, o RStudio.

O R é um software especializado em manipulação, análise e visualização gráfica de dados estatísticos que utiliza uma linguagem

de programação homônima. Além disso, é um software expansível graças à possibilidade de utilização dos *packages* (cf. GRIES, 2013) com dados e funções para diferentes áreas do conhecimento relacionado à estatística sendo, atualmente, considerado um dos melhores ambientes computacionais para o tratamento de dados estatísticos.

O RStudio é uma interface funcional e mais amigável para o R, sendo o principal ambiente de desenvolvimento integrado para R, que disponibiliza ferramentas adicionais diretamente na interface gráfica, tais como a visualização dos *scripts* abertos recentemente, o histórico de linhas de comando executadas, a lista de pacotes instalados, entre outras (OUSHIRO, 2014).

Além de executar vários modelos estatísticos, como a regressão linear, e os demais modelos da família de modelos lineares generalizados (Regressão Logística, Ordinal, Poisson, Multinomial, etc.), o R analisa, também, variáveis previsoras de efeitos aleatórios por meio de modelos mistos, executando, ainda, diversas outras tarefas, como elaboração de gráficos diversos, tabelas e cálculos de medidas estatísticas, tais como, média, mediana, desvio padrão, variância etc. (cf. GRIES, 2013; LEVSHINA, 2015; OUSHIRO, 2017; WINTER, 2020).