**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Modelos estatísticos para dados politômicos nominais em estudos
longitudinais com uma aplicação à área agronômica**

Vinícius Menarin

Dissertação apresentada para obtenção do título de
Mestre em Ciências. Área de concentração: Estatística
e Experimentação Agronômica

**Piracicaba
2016**

Vinícius Menarin
Bacharel em Estatística

Modelos estatísticos para dados politômicos nominais em estudos longitudinais com uma aplicação à área agronômica

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **IDEMAURO ANTONIO RODRIGUES DE LARA**

Dissertação apresentada para obtenção do título de Mestre em Ciências. Área de concentração: Estatística e Experimentação Agronômica

**Piracicaba
2016**

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Menarin, Vinícius

Modelos estatísticos para dados politômicos nominais em estudos longitudinais com uma aplicação à área agrônômica / Vinícius Menarin. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2016.

106 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz".

1. Medidas repetidas no tempo 2. Dados categorizados nominais 3. Equações de estimação generalizadas 4. Modelos lineares generalizados mistos I. Título

CDD 630.219
M535m

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

Dedicatória

Dedico este trabalho às seguintes pessoas, de fundamental importância para mim, que não poderia deixar de lembrar:

*Aos meus pais, **Luiz e Lúcia**, meus exemplos de vida íntegra e honesta. Muito obrigado pelos ensinamentos, pelo apoio incondicional, pelos valores e educação transmitidos até os dias de hoje. Sem o apoio de vocês não teria alcançado esta meta.*

*Ao meu irmão **Henrique**, pelas ideias, experiências trocadas e pelo incentivo constante.*

*À minha avó, **Dalva**, que infelizmente não pôde acompanhar até o fim esta conquista mas que sempre torceu por mim e ficaria muito feliz em vê-la.*

*À minha namorada **Giane**, que esteve ao meu lado com paciência e compreensão nos momentos difíceis, sempre demonstrando seu carinho e afeto ao longo dos dois anos de convivência à distância.*

AGRADECIMENTOS

A Deus, que iluminando meus caminhos permitiu-me alcançar esta meta pessoal e profissional dando-me força e sabedoria indispensáveis.

Ao professor doutor Idemauro Antonio Rodrigues de Lara, meu orientador, com quem tive o privilégio de cursar disciplinas na graduação e no mestrado. Obrigado por aceitar o projeto, pelos ensinamentos, conselhos e por acreditar em mim e na conclusão deste trabalho.

Aos membros da banca pela avaliação do trabalho, sugestões e contribuições dadas.

Ao professor doutor Sila Carneiro da Silva, do Departamento de Zootecnia da ESALQ, e ao colega Ricardo Klein pela disponibilização do conjunto de dados utilizado neste trabalho.

Aos pesquisadores Anestis Touloumis, John M. Williamson e KyungMann Kim, autores de programas computacionais desenvolvidos na área deste trabalho, por responderem aos contatos feitos e dúvidas esclarecidas.

Aos colegas da pós-graduação, em especial Viviane Panariello Paulenas, Maíra Blumer Faretto e Christopher Silva de Pádua pela convivência, amizade, trabalhos desenvolvidos em parceria e experiências trocadas nesse período.

Aos professores do programa de pós-graduação em Estatística e Experimentação Agronômica da ESALQ com quem tive a satisfação de cursar disciplinas. O conhecimento adquirido nas disciplinas será de grande contribuição para minha vida profissional.

Aos professores da graduação em Estatística pela UFPR, responsáveis pela base teórica indispensável no mestrado. Obrigado também àqueles que me estimularam a cursar o mestrado quando comentei que tinha essa intenção, professores Suely Ruiz Giolo e Cesar Augusto Taconelli.

Aos funcionários do Departamento de Ciências Exatas e do programa da pós-graduação, pelo suporte dado no decorrer do curso.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), fundação do Ministério da Educação (MEC), pelo importante apoio financeiro durante o curso.

“Um cientista deve acima de tudo ser como uma criança. Se ele vê algo, deve dizer o que está vendo, independente daquilo ser o que ele imaginava ver. Ver primeiro, testar depois. Mas sempre ver primeiro. Senão você só vai ver o que espera ver.”

Douglas Adams

“Não podemos escapar dos dados, assim como não podemos evitar o uso de palavras. Tais como palavras os dados não se interpretam a si mesmos, mas devem ser lidos com entendimento. Da mesma maneira que um escritor pode dispor as palavras em argumentos convincentes ou frases sem sentido, assim também os dados podem ser convincentes, enganosos ou simplesmente inócuos. A instrução numérica, a capacidade de acompanhar e compreender argumentos baseados em dados, é importante para qualquer um de nós. O estudo da estatística é parte essencial de uma formação sólida.”

Moore

SUMÁRIO

RESUMO	11
ABSTRACT	13
LISTA DE ILUSTRAÇÕES	15
LISTA DE TABELAS	17
1 INTRODUÇÃO	19
2 REVISÃO BIBLIOGRÁFICA	21
2.1 Dados categorizados	21
2.1.1 Modelos lineares generalizados	21
2.1.1.1 Família exponencial de distribuições	22
2.1.1.2 Definição	22
2.1.1.3 Estimação dos parâmetros	23
2.1.1.4 Seleção de modelos	24
2.1.2 Modelos para dados categorizados	25
2.1.2.1 Variáveis dicotômicas	26
2.1.2.2 Variáveis politômicas	27
2.2 Estudos longitudinais	29
2.2.1 Modelos marginais	31
2.2.1.1 Equações de estimação generalizadas (EEG)	31
2.2.2 Modelos de efeitos aleatórios	35
2.2.3 Diferenças entre modelos marginais e modelos de efeitos aleatórios	39
2.3 Extensões para respostas politômicas nominais longitudinais	43
2.3.1 Extensão das EEG para o caso politômico	43
2.3.1.1 Metodologia	43
2.3.1.2 Recursos computacionais	45
2.3.2 EEG utilizando uma parametrização de razões de chances locais	45
2.3.2.1 Metodologia	45
2.3.2.2 Recursos computacionais	48
2.3.3 Modelos mistos	48
2.3.3.1 Metodologia	49
2.3.3.2 Recursos computacionais	49
3 MATERIAL E MÉTODOS	51
3.1 Material	51
3.2 Métodos	52

3.2.1	Ajuste via EEG	52
3.2.1.1	Seleção da estrutura de associação e das variáveis	52
3.2.1.2	Probabilidades preditas e intervalos de confiança	55
3.2.1.3	Avaliação da qualidade do ajuste	58
3.2.1.4	Interpretações dos resultados	58
3.2.2	Ajuste via modelos mistos	58
3.2.2.1	Seleção do modelo	58
3.2.2.2	Probabilidades preditas e intervalos de confiança	59
3.2.2.3	Avaliação da qualidade do ajuste	60
3.2.2.4	Interpretações dos resultados	60
4	RESULTADOS E DISCUSSÃO	61
4.1	Análise exploratória	61
4.2	Modelo ajustado via EEG	62
4.2.1	Seleção e formulação do modelo	62
4.2.2	Qualidade do ajuste	66
4.2.3	Interpretações e discussão	68
4.3	Modelo misto	76
4.3.1	Seleção e formulação do modelo	76
4.3.2	Qualidade do ajuste	78
4.3.3	Interpretações e discussão	79
5	CONSIDERAÇÕES FINAIS	83
	REFERÊNCIAS	85
	APÊNDICES	89
	ANEXOS	93

RESUMO

Modelos estatísticos para dados politômicos nominais em estudos longitudinais com uma aplicação à área agrônômica

Estudos em que a resposta de interesse é uma variável categorizada são bastante comuns nas mais diversas áreas da Ciência. Em muitas situações essa resposta é composta por mais de duas categorias não ordenadas, denominada então de uma variável politômica nominal, e em geral o objetivo do estudo é associar a probabilidade de ocorrência de cada categoria aos efeitos de variáveis explicativas. Ademais, existem tipos especiais de estudos em que os dados são coletados diversas vezes para uma mesma unidade amostral ao longo do tempo, os estudos longitudinais. Estudos assim requerem o uso de modelos estatísticos que considerem em sua formulação algum tipo de estrutura que suporte a dependência que tende a surgir entre observações feitas em uma mesma unidade amostral. Neste trabalho são abordadas duas extensões do modelo de logitos generalizados, usualmente empregado quando a resposta é politômica nominal com observações independentes entre si. A primeira consiste de uma modificação das equações de estimação generalizadas para dados nominais que se utiliza de razões de chances locais para descrever a dependência entre as observações da variável resposta politômica ao longo dos diversos tempos observados. Este tipo de modelo é denominado de modelo marginal. A segunda proposta abordada consiste no modelo de logitos generalizados com a inclusão de efeitos aleatórios no preditor linear, que também leva em conta uma dependência entre as observações. Esta abordagem caracteriza o modelo de logitos generalizados misto. Há diferenças importantes inerentes às interpretações dos modelos marginais e mistos, que são discutidas e que devem ser levadas em consideração na escolha da abordagem adequada. Ambas as propostas são aplicadas em um conjunto de dados proveniente de um experimento da área agrônômica realizado em campo, conduzido sob um delineamento casualizado em blocos com esquema fatorial para os tratamentos. O experimento foi acompanhado ao longo de seis estações do ano, caracterizando assim uma estrutura longitudinal, sendo a variável resposta o tipo de vegetação observado no campo (touceiras, plantas invasoras ou espaços vazios). Os resultados encontrados são satisfatórios, embora a dependência presente nos dados não seja tão caracterizada; por meio de testes como da razão de verossimilhanças e de Wald diversas diferenças significativas entre os tratamentos foram encontradas. Ainda, devido às diferenças metodológicas das duas abordagens, o modelo marginal baseado nas equações de estimação generalizadas mostra-se mais adequado para esses dados.

Palavras-chave: Medidas repetidas no tempo; Dados categorizados nominais; Equações de estimação generalizadas; Modelos lineares generalizados mistos

ABSTRACT

Statistical models for nominal polytomous data in longitudinal studies with an application to agronomy

Studies where the response is a categorical variable are quite common in many fields of Sciences. In many situations this response is composed by more than two unordered categories characterizing a nominal polytomous outcome and, in general, the aim of the study is to associate the probability of occurrence of each category to the effects of variables. Furthermore, there are special types of study where many measurements are taken over the time for the same sampling unit, called longitudinal studies. Such studies require special statistical models that consider some kind of structure that support the dependence that tends to arise from the repeated measurements for the same sampling unit. This work focuses on two extensions of the baseline-category logit model usually employed in cases when there is a nominal polytomous response with independent observations. The first one consists in a modification of the well-known generalized estimating equations for longitudinal data based on local odds ratios to describe the dependence between the levels of the response over the repeated measurements. This type of model is also known as a marginal model. The second approach adds random effects to the linear predictor of the baseline-category logit model, which also considers a dependence between the observations. This characterizes a baseline-category mixed model. There are substantial differences inherent to interpretations when marginal and mixed models are compared, what should be considered in the choice of the most appropriated approach for each situation. Both methodologies are applied to the data of an agronomic experiment installed under a complete randomized block design with a factorial arrangement for the treatments. It was carried out over six seasons, characterizing the longitudinal structure, and the response is the type of vegetation observed in field (tussocks, weeds or regions with bare ground). The results are satisfactory, even if the dependence found in data is not so strong, and likelihood-ratio and Wald tests point to several differences between treatments. Moreover, due to methodological differences between the two approaches, the marginal model based on generalized estimating equations seems to be more appropriate for this data.

Keywords: Repeated measurements over time; Nominal categorical data; Generalized estimating equations; Generalized linear mixed models

LISTA DE ILUSTRAÇÕES

Figura 2.1 – Representação de um modelo com interceptos aleatórios	41
Figura 2.2 – Comparação de um modelo marginal e de um modelo misto: (a) efeito marginal obtido em um MLGM, (b) efeito marginal obtido por um ajuste via EEG e (c) efeito observado para um indivíduo com efeito aleatório nulo . .	42
Figura 3.1 – Croqui do experimento realizado	51
Figura 4.1 – Proporções médias observadas de cada tipo de vegetação ao longo das estações do ano, conforme os tratamentos utilizados	62
Figura 4.2 – Comparação das proporções observadas e previstas pelo modelo 6, ajustado via EEG	67
Figura 4.3 – Correlação entre os valores observados e previstos pelo modelo 6, ajustado via EEG	67
Figura 4.4 – Resíduos do modelo 6 ajustado via EEG, de acordo com cada um dos logitos	68
Figura 4.5 – Intervalos de 95% de confiança para as probabilidades previstas pelo modelo 6, ajustado via EEG	70
Figura 4.6 – Comparação das proporções observadas e previstas pelo modelo misto ajustado, com a inclusão de efeitos aleatórios no modelo 6	78
Figura 4.7 – Correlação entre os valores observados e previstos pelo modelo misto ajustado, com a inclusão de efeitos aleatórios no modelo 6	79
Figura 4.8 – Probabilidades individuais e médias estimadas, considerando a presença de efeitos aleatórios ao preditor linear do modelo 6	80
Figura 4.9 – Intervalos de 95% de confiança para as probabilidades previstas pelo modelo misto, com a inclusão de efeitos aleatórios no modelo 6	80

LISTA DE TABELAS

Tabela 2.1 – Estrutura de um conjunto de dados longitudinais com resposta contínua . . .	30
Tabela 2.2 – Exemplo de tabela de contingência formada no cruzamento de dois instantes de tempo	46
Tabela 3.1 – Quantidade de observações analisadas no experimento, por tratamento e estação do ano	52
Tabela 3.2 – Modelos encaixados propostos para o ajuste via EEG	53
Tabela 3.3 – Codificações das variáveis <i>dummy</i> utilizadas no modelo ajustado via EEG .	55
Tabela 4.1 – Proporções médias observadas (p) de touceira, planta invasora e espaço vazio em cada tratamento e estação do ano, e seus respectivos desvios padrões (d.p.)	61
Tabela 4.2 – Estimativas dos parâmetros intrínsecos do ajuste via EEG, utilizando a estrutura RC	63
Tabela 4.3 – Resultado do teste de Wald nas comparações dos modelos encaixados ajustados via EEG	64
Tabela 4.4 – Estimativas dos parâmetros e erros padrões (entre parênteses) do modelo 6, ajustado via EEG	65
Tabela 4.5 – Estimativas dos parâmetros e erros padrões (entre parênteses) do modelo ajustado 6 via EEG, para a comparação touceira $\times$ planta invasora	66
Tabela 4.6 – Estimativas das razões de chances ($\hat{\psi}$) entre os níveis das variáveis do modelo 6 ajustado via EEG, e respectivos intervalos de confiança (IC)	71
Tabela 4.7 – Erros padrões (EP) <i>naïve</i> e robustos das estimativas dos parâmetros referentes ao modelo 6, ajustado via EEG	75
Tabela 4.8 – Estimativas dos parâmetros e erros padrões (entre parênteses) do modelo misto ajustado, com a inclusão de efeitos aleatórios no modelo 6	77
Tabela 4.9 – Estimativas dos componentes de variância do modelo misto ajustado, com a inclusão de efeitos aleatórios no modelo 6	78
Tabela 4.10–Estatísticas de ajuste do modelo misto com a inclusão de efeitos aleatórios no modelo 6 e de um modelo sem efeitos aleatórios	78
Tabela 4.11–Estimativas das razões de chances ($\hat{\psi}$) entre as estações do ano de acordo com o modelo misto ajustado com base no modelo 6, com respectivos intervalos de confiança (IC)	81

1 INTRODUÇÃO

Estudos científicos, das mais diversas áreas, são comumente realizados de tal forma que observações da característica de interesse (variável resposta) são coletadas de uma mesma unidade amostral (indivíduo, em muitos casos) diversas vezes. Exemplos ocorrem quando as mensurações são feitas ao longo do tempo, em diferentes doses, distâncias, entre outras, sendo designados muitas vezes pelo abrangente termo medidas repetidas. Todavia, a primeira situação exemplificada, quando isso se dá ao longo do tempo, é mais frequentemente denominada por estudos longitudinais ou ainda “em painel” em algumas áreas. Caso contrário, quando somente uma observação é feita por unidade amostral diz-se que o estudo é transversal.

Um modelo estatístico, o processo de explicar o comportamento da variável resposta de acordo com variáveis explicativas estatisticamente relevantes por meio de uma expressão matemática, depende essencialmente se o estudo conduzido é transversal ou longitudinal. Em linhas gerais, nos estudos transversais utilizam-se modelos clássicos em que se assumem observações independentes entre si (uma vez que os indivíduos são independentes e tem-se somente uma observação para cada um); porém, no cenário longitudinal, naturalmente espera-se que observações feitas em um mesmo indivíduo sejam dependentes entre si e, portanto, devem ser utilizados modelos especiais que levem isso em consideração na sua formulação. Estes modelos podem ser de três tipos, essencialmente: marginais, que utilizam as equações de estimação generalizadas (EEG) (LIANG; ZEGGER, 1986; ZEGGER; LIANG, 1992), mistos (de efeitos aleatórios) e de transição (DIGGLE *et al.*, 2002).

Paralelamente a essas definições, distinguem-se ainda os tipos de variáveis respostas em dois grandes grupos: qualitativas e quantitativas. As qualitativas, objeto de estudo deste trabalho, ocorrem quando a resposta de interesse não pode ser medida em uma escala numérica e atribuem-se categorias a ela. Na Estatística, variáveis deste tipo são estudadas pela área de dados categorizados, também conhecidos por categóricos ou discretos, esta de relevante importância em diversos campos do conhecimento humano como as ciências biológicas, sociais, educacionais, exatas e humanas, por exemplo. Distinguem-se ainda, na ótica das variáveis categorizadas, as dicotômicas (ou binárias, quando constituídas por duas categorias) e politômicas, no caso de haver três ou mais categorias. Ademais, recebem a classificação entre ordinais e nominais, quando há uma ordenação natural entre as categorias ou não, respectivamente. Em alguns casos, quando a variável é politômica nominal ela é denominada também de multinomial (AGRESTI, 2002; AGRESTI, 2007).

Em se tratando de estudos longitudinais existem metodologias desenvolvidas para o caso politômico ordinal (HEAGERTY; ZEGGER, 1996; HEDEKER; GIBBONS, 1994; MILLER; DAVIS; LANDIS, 1993), utilizadas em trabalhos como Santana (2004) e Silva (2013). No

entanto, apesar da relativa importância das variáveis nominais, verificou-se, tanto do ponto de vista computacional quanto metodológico, uma necessidade de abordar este caso com maior ênfase, que serviu de motivação inicial para este trabalho. Dentre as três classes de modelos citadas abordam-se aqui os modelos marginais utilizando EEG e os modelos de efeitos aleatórios, e respectivas implementações nos *softwares* R (R Core Team, 2014) e SAS (SAS, 2011), versão 9.3. Há uma carência até o presente momento de recursos computacionais capazes de ajustar modelos na estrutura estudada neste trabalho, devido ao fato que a grande parte dos recursos concentra-se no cenário ordinal.

Os objetivos deste trabalho são, portanto:

- i) fazer uma revisão das metodologias propostas para a análise de dados politômicos nominais longitudinais, no contexto das equações de estimação generalizadas e dos modelos mistos;
- ii) apresentar as diferenças entre ambas as abordagens focando principalmente nas interpretações para cada tipo de modelo;
- iii) por fim, desenvolver uma aplicação em um conjunto de dados referente a um experimento realizado em campo, na área agrônômica, com a construção de intervalos de confiança para as probabilidades das categorias de resposta. Na literatura consultada nem sempre é demonstrado como obtê-los, são encontrados apenas exemplos de modelos simples com poucas variáveis ou somente para o caso dicotômico.

O texto desta dissertação está organizado na seguinte ordem: inicialmente são apresentados os modelos desenvolvidos para dados categorizados, com estrutura transversal. Posteriormente, os modelos para dados longitudinais (com qualquer tipo de resposta) baseados nas equações de estimação generalizadas e nos modelos mistos. Na sequência, apresentam-se os modelos que unificam as abordagens apresentadas até então, ou seja, aqueles desenvolvidos para dados politômicos nominais longitudinais. Por fim, é mostrada a aplicação de tais modelos a um exemplo e são discutidos os principais resultados e conclusões deste trabalho.

2 REVISÃO BIBLIOGRÁFICA

2.1 Dados categorizados

Quando uma variável tem sua escala de mensuração formada por um conjunto de categorias, diz-se que ela é uma variável categorizada. Os métodos de análise para variáveis deste tipo levaram mais tempo para serem desenvolvidos, quando comparados àqueles para variáveis quantitativas. Um dos pioneiros no desenvolvimento de trabalhos iniciais foi o estatístico britânico Karl Pearson, por volta de 1900. Contudo, houve pouco desenvolvimento desta área até a década de 1960 quando, então, estudos nas áreas de ciências sociais e biomédicas impulsionaram o desenvolvimento de tais métodos. Agresti (2002) traça um histórico do desenvolvimento dos métodos para a análise de dados categorizados, além de ser uma referência clássica para estudos nesta área.

Em se tratando da classificação de variáveis categorizadas, primeiramente elas podem ser classificadas quanto ao número de categorias de que são compostas: dicotômicas (duas categorias) ou politômicas (três ou mais categorias). Outra classificação necessária é quanto à natureza destas categorias: ordinais, quando apresentam uma ordenação natural, ou nominais no caso de não haver essa ordenação (GIOLO, 2012). São exemplos:

- i) Nominais:** classificação de uma planta em sadia ou doente, pesticida adotado no combate a uma peste no campo (A, B ou C), tipo de ordenha de um rebanho leiteiro (manual ou mecanizada);
- ii) Ordinais:** estágio de infestação de uma plantação por pragas (inicial, intermediário ou avançado), grau de lesões em animais (baixo, médio ou alto).

Em termos metodológicos, há abordagens e modelos estatísticos específicos para a análise de variáveis nominais e ordinais, devendo-se tomar o devido cuidado na escolha daquele a ser empregado. Agresti (2007) ressalta que métodos desenvolvidos para o caso ordinal não devem ser aplicados ao caso nominal e que, embora haja uma perda relevante de informação, métodos desenvolvidos para o caso nominal podem ser aplicados ao ordinal.

A classe de modelos lineares generalizados, apresentada a seguir, é a base para vários modelos que podem ser utilizados para analisar dados cuja variável resposta seja categorizada.

2.1.1 Modelos lineares generalizados

Nelder e Wedderburn (1972) introduziram a teoria de modelos lineares generalizados (MLG), unificando abordagens que até então haviam sido estudadas separadamente para diversos tipos de dados. Os MLG permitem analisar dados de natureza discreta ou contínua e constituem uma ferramenta poderosa de análise de dados.

2.1.1.1 Família exponencial de distribuições

Antes de propriamente definir os modelos lineares generalizados (MLG) é preciso detalhar alguns aspectos da família exponencial de distribuições. Uma variável aleatória (v.a.) uniparamétrica Y é dita pertencer a esta família se sua função (densidade ou de probabilidade), $f(y; \theta)$, puder ser escrita na forma

$$f(y; \theta) = h(y) \exp [\eta(\theta)t(y) - b(\theta)], \quad (2.1)$$

que na sua forma canônica (considerando $\eta(\theta)$ e $t(y)$ iguais à função identidade) fica expressa por

$$f(y; \theta) = h(y) \exp [\theta y - b(\theta)], \quad (2.2)$$

em que θ é chamado de parâmetro canônico. Nelder e Wedderburn (1972) formularam a teoria dos MLG partindo da expressão (2.2), incorporando ainda um parâmetro adicional, $\phi > 0$, que representa uma medida de dispersão. Com isso, tem-se

$$f(y; \theta, \phi) = \exp \left\{ \phi^{-1} [y\theta - b(\theta)] + c(y, \phi) \right\}, \quad (2.3)$$

sendo as funções $b(\cdot)$ e $c(\cdot)$ conhecidas. Pode ser demonstrado, por meio das propriedades da função escore, que a esperança e a variância de uma v.a. Y pertencente à família exponencial (2.3) são dadas por $E[Y] = \mu = b'(\theta)$ e $\text{Var}[Y] = \phi b''(\theta) = \phi V(\mu)$, em que $V(\mu)$ é chamada de função de variância.

Há ainda uma extensão da família exponencial uniparamétrica (2.1), denominada de família exponencial multiparamétrica, de dimensão k , caracterizada por uma função (densidade ou de probabilidade) da forma

$$f(\mathbf{y}; \boldsymbol{\theta}) = h(\mathbf{y}) \exp \left[\sum_{i=1}^k \eta_i(\boldsymbol{\theta}) t_i(\mathbf{y}) - b(\boldsymbol{\theta}) \right],$$

sendo $\boldsymbol{\theta}$ um vetor de parâmetros de dimensão k e $\eta_i(\boldsymbol{\theta})$, $b(\boldsymbol{\theta})$, $t_i(\mathbf{y})$ e $h(\mathbf{y})$ funções definidas em subconjuntos dos reais. Quando $\eta_i(\boldsymbol{\theta}) = \theta_i$ tem-se a família exponencial na forma canônica

$$f(\mathbf{y}; \boldsymbol{\theta}) = h(\mathbf{y}) \exp \left[\sum_{i=1}^k \theta_i t_i(\mathbf{y}) - b(\boldsymbol{\theta}) \right]$$

com parâmetros canônicos $\theta_1, \dots, \theta_k$. Algumas distribuições pertencentes à família exponencial multiparamétrica são: normal, normal inversa, multinomial, gama e beta. As distribuições binomial e Poisson são exemplos do caso uniparamétrico (DEMÉTRIO; CORDEIRO; MORAL, 2014).

2.1.1.2 Definição

Um MLG é composto por três partes (DEMÉTRIO; CORDEIRO; MORAL, 2014):

i) Componente aleatório: seja um conjunto de n v.a. independentes e identicamente distribuídas (i.i.d.) $Y_1, Y_2, \dots, Y_n$, com distribuição pertencente à família (2.3). A função (densidade ou de probabilidade) é escrita como:

$$f(y_i; \theta_i, \phi) = \exp \left\{ \phi^{-1} [y_i \theta_i - b(\theta_i)] + c(y_i, \phi) \right\} \quad i = 1, \dots, n. \quad (2.4)$$

Analogamente ao exposto na seção 2.1.1.1, $E[Y_i] = \mu_i = b'(\theta_i)$ e $\text{Var}[Y_i] = \phi b''(\theta_i) = \phi V_i = \phi V(\mu_i)$.

ii) Componente sistemático: são as covariáveis (ou variáveis explicativas, regressoras, independentes), que entram como uma estrutura linear no modelo (preditor linear). Este componente é estabelecido no planejamento do experimento, e pode resultar em modelos de regressão linear simples ou múltipla, de análise de variância (delineamentos inteiramente casualizados, casualizados em blocos, quadrados latinos, etc.), ou análise de covariância (cujas covariáveis são fatores e variáveis contínuas). O preditor linear é representado por

$$\eta_i = \sum_{r=1}^p x_{ir} \beta_r = \mathbf{x}_i' \boldsymbol{\beta} \quad \text{ou} \quad \boldsymbol{\eta} = \mathbf{X} \boldsymbol{\beta}$$

em que $\boldsymbol{\beta}$ é um vetor de p parâmetros desconhecidos e $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]'$ é a matriz de delineamento ($n \times p$);

iii) Função de ligação: é uma função monótona e diferenciável, representada por $g(\mu_i)$, que relaciona os componentes sistemático e aleatório vinculando a média da distribuição da variável resposta ao preditor linear (covariáveis) da seguinte forma:

$$\eta_i = g(\mu_i).$$

Se a função de ligação for determinada de forma que $g(\mu_i) = \theta_i = \eta_i$ ela é chamada de função de ligação canônica pois o preditor linear η_i modela diretamente o parâmetro canônico θ_i . Dessa forma, cada distribuição de probabilidades possui sua função de ligação canônica.

2.1.1.3 Estimação dos parâmetros

Dentre vários métodos propostos para a estimação do vetor de parâmetros $\boldsymbol{\beta}$ o mais utilizado é de máxima verossimilhança (MV), que tem propriedades de interesse como consistência e eficiência assintótica. Da expressão (2.4), o logaritmo da função de verossimilhança (considerando ϕ conhecido) é

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n l_i(\theta_i, \phi; y_i) = \phi^{-1} \sum_{i=1}^n [y_i \theta_i - b(\theta_i)] + \sum_{i=1}^n c(y_i, \phi),$$

em que $\theta_i = q(\mu_i)$, $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \sum_{r=1}^p x_{ir} \beta_r$. Os parâmetros $\theta_1, \dots, \theta_n$ e $\mu_1, \dots, \mu_n$ são desconhecidos, mas são também funções dos parâmetros lineares $\beta_1, \dots, \beta_p$ do modelo. Assim

sendo, estimando-se o vetor β consegue-se depois chegar às estimativas dos vetores θ e μ por meio das relações $\theta_i = q(\mu_i)$ e $\mu_i = g^{-1}(x_i'\beta)$.

Aplicando-se o método da máxima verossimilhança e fazendo uso da regra da cadeia, tem-se que o vetor escore $U(\beta)$ tem seus elementos dados por

$$U_r = \frac{\partial l(\beta)}{\partial \beta_r} = \sum_{i=1}^n \frac{dl_i}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_r} \quad r = 1, \dots, p. \quad (2.5)$$

Sabe-se que $\mu_i = b'(\theta_i)$ e $V_i = \frac{d\mu_i}{d\theta_i}$, então a expressão 2.5 fica reescrita como

$$U_r = \phi^{-1} \sum_{i=1}^n (y_i - \mu_i) \frac{1}{V_i} \frac{d\mu_i}{d\eta_i} x_{ir} \quad r = 1, \dots, p.$$

A estimação dos parâmetros β por máxima verossimilhança consiste em igualar as p equações U_r a zero. Entretanto, tais equações não são lineares e são precisos métodos iterativos para sua solução. Resumidamente, o processo desenvolvido por Nelder e Wedderburn (1972) e discutido em detalhes por Demétrio, Cordeiro e Moral (2014) se dá por mínimos quadrados ponderados iterativamente, com a solução sendo dada por

$$\beta^{(m+1)} = (X'W^{(m)}X)^{-1} X'W^{(m)}z^{(m)},$$

em que m denota cada passo do processo iterativo, W é uma matriz de pesos e $z_i^{(m)}$ é dado por $z_i^{(m)} = \eta_i^{(m)} + (y_i - \mu_i^{(m)})g'(\mu_i^{(m)})$.

2.1.1.4 Seleção de modelos

No processo de seleção de modelos deve-se buscar uma estrutura final que explique os dados corretamente e com parcimônia, ou seja, que seja a mais simples possível (menor número de parâmetros). Ajustar um modelo significa substituir os valores observados (y) por valores estimados ($\hat{\mu}$) de acordo com as covariáveis (x), de forma que as diferenças observadas entre eles sejam consideradas toleráveis segundo algum critério.

No tocante à parcimônia, um modelo com muitas covariáveis (e consequentemente muitos parâmetros) tem um bom ajuste aos dados, mas as interpretações acerca dos parâmetros são mais complexas. Por outro lado, um modelo com poucas covariáveis tem interpretações mais simples mas com um ajuste pior aos dados. O desafio está em encontrar um modelo satisfatório entre estes dois extremos, representados comumente por dois modelos: nulo e saturado.

O primeiro, modelo nulo, é o mais simples possível contendo apenas um parâmetro que representa um valor comum a todas as observações; neste modelo, toda a variabilidade existente entre as respostas y fica atribuída ao componente aleatório do modelo. Em compensação, o modelo saturado contém o máximo de parâmetros possíveis (n): há um parâmetro para cada observação. Naturalmente, a qualidade do ajuste é perfeita pois toda a variabilidade é atribuída

ao componente sistemático. Porém, esses dois extremos não são de grande valia na prática por resumir demais ou deixar de explicar a variabilidade total (LARA, 2007; DEMÉTRIO; CORDEIRO; MORAL, 2014).

Nelder e Wedderburn (1972) definiram uma medida de discrepância para comparar um modelo em estudo com o modelo saturado, a função desvio (*deviance*):

$$S_p = 2 \left(\hat{l}_n - \hat{l}_p \right), \quad (2.6)$$

em que $\hat{l}_n$ e $\hat{l}_p$ representam os máximos dos logaritmos das funções de verossimilhança para os modelos saturado e sob estudo, respectivamente. Tem-se ainda que

$$S_p = \phi^{-1} D_p,$$

sendo D_p denominado somente de desvio e S_p de desvio escalonado devido à presença do parâmetro ϕ . O valor de D_p é sempre positivo e decresce à medida que mais variáveis são adicionadas ao modelo; logo, quanto melhor o ajuste do modelo aos dados menor é o valor de D_p e por isso é utilizado como um indicativo da qualidade do ajuste.

Com isso, quando da seleção de modelos para MLG, o processo conhecido por ANODEV (*analysis of deviance* - análise do desvio) é empregado. Este método pode ser visto como uma generalização da ANOVA (*analysis of variance* - análise da variância), empregada para os modelos de regressão convencionais. Sejam M_p e M_q dois modelos encaixados com p e q parâmetros de modo que $p < q$. A estatística $D_p - D_q$ é interpretada como a quantidade da variabilidade dos dados que é explicada pelos termos que estão presentes no modelo M_q mas que não estão em M_p (incluídos os efeitos de M_p). Quando o parâmetro ϕ é conhecido (distribuições binomial e Poisson, por exemplo), assintoticamente tem-se que

$$S_q - S_p = \phi^{-1} (D_q - D_p) \sim \chi^2_{p-q},$$

coincidindo neste caso com o teste da razão de verossimilhanças (TRV). Se o resultado deste teste for significativo adota-se o modelo com maior número de parâmetros M_p , pois a redução no desvio foi significativa. Para casos em que o parâmetro de dispersão (ϕ) é desconhecido ele precisa ser estimado de forma consistente (preferencialmente utilizando-se o modelo maximal com m parâmetros), e a estatística a ser utilizada é

$$F = \frac{(D_q - D_p)/(p - q)}{\hat{\phi}} \sim F_{p-q, n-m}.$$

2.1.2 Modelos para dados categorizados

A seguir são apresentados os modelos comumente aplicados à análise de dados categorizados nominais, sejam eles dicotômicos ou politômicos.

2.1.2.1 Variáveis dicotômicas

Para variáveis categorizadas dicotômicas, geralmente, assume-se que os dados resultam de n ensaios independentes e idênticos com duas possíveis respostas (classificações), frequentemente chamadas de “sucesso” e “fracasso”. Assumir ensaios idênticos significa que a probabilidade de sucesso é a mesma para todos, e por ensaios independentes entende-se que as respostas são variáveis aleatórias independentes, ou ainda que a resposta de um ensaio não interfere em outro. Na literatura isso também é conhecido por *ensaios de Bernoulli*. Feitas tais considerações, Y tem distribuição binomial com probabilidade de sucesso em cada ensaio π . A probabilidade de ocorrência de y sucessos em n ensaios é dada por

$$P[Y = y] = \frac{n!}{y!(n-y)!} \pi^y (1-\pi)^{n-y}, y = 0, 1, \dots, n.$$

A média e a variância da distribuição binomial são dadas respectivamente por $E[Y] = \mu = n\pi$ e $\text{Var}[Y] = n\pi(1-\pi)$. Pelo que foi apresentado na seção 2.1.1.2, a função de ligação canônica da distribuição binomial é dada por

$$\eta = \ln \left(\frac{\pi}{1-\pi} \right) = \ln \left(\frac{\mu}{n-\mu} \right).$$

Esta função de ligação é chamada de logito e caracteriza o que é conhecido por modelo logístico. Denotando o preditor linear na forma $\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = \mathbf{x}'\boldsymbol{\beta}$, o modelo logístico (em função das covariáveis $\mathbf{x}$) fica definido pela relação:

$$\ln \left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} \right) = \mathbf{x}'\boldsymbol{\beta}.$$

A probabilidade de ocorrência do evento escolhido como “sucesso” é dada por

$$\pi(\mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}. \quad (2.7)$$

Considerando, por exemplo, um preditor linear da forma $\eta = \beta_0 + \beta_1 x$, se $\beta_1 > 0$ tem-se que a probabilidade π aumenta à medida que x também aumenta. Se $\beta_1 < 0$, à medida que x aumenta a probabilidade π diminui. Outra forma bastante intuitiva e utilizada de interpretar os parâmetros estimados em um modelo logístico é por meio das razões de chances (*odds ratios*), simbolizadas por ψ . Para o modelo simples exemplificado neste parágrafo, a chance (*odds*) de sucesso é dada por

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\beta_0 + \beta_1 x) = e^{\beta_0} (e^{\beta_1})^x.$$

Esta expressão permite concluir que, para cada unidade de acréscimo em x , a chance de sucesso é multiplicada por e^{β_1} pois a razão de chances (ψ) é dada por

$$\psi = \frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}} = \frac{e^{\beta_0} (e^{\beta_1})^{x+1}}{e^{\beta_0} (e^{\beta_1})^x} = e^{\beta_1}, \quad (2.8)$$

ou seja, a chance de sucesso de um indivíduo com valor $(x + 1)$ é (e^{β_1}) vezes a chance de sucesso de um indivíduo com valor x . Quando x for uma variável qualitativa, (e^{β_1}) representa o aumento na chance de sucesso quando se comparam os níveis desta variável (sempre em relação ao nível de referência) (AGRESTI, 2007). Quando há mais variáveis explicativas no preditor linear a interpretação dos parâmetros é análoga, sempre fixando os níveis das variáveis em que não há interesse em comparar.

Pelas propriedades assintóticas dos estimadores, assumindo-se que têm distribuição normal, intervalos de confiança podem ser obtidos tanto para as probabilidades preditas quanto para as razões de chances. Considerando novamente, a título de exemplo, o preditor linear $\eta = \beta_0 + \beta_1 x$, o valor estimado do logito é

$$\text{logito} [\hat{\pi}(x)] = \hat{\beta}_0 + \hat{\beta}_1 x,$$

cuja variância estimada é

$$\hat{\text{Var}} [\text{logito} [\hat{\pi}(x)]] = \hat{\text{Var}} [\hat{\beta}_0] + \hat{\text{Var}} [\hat{\beta}_1] + 2\hat{\text{cov}} [\hat{\beta}_0, \hat{\beta}_1].$$

Portanto, o intervalo de confiança para o logito é

$$\text{logito} [\hat{\pi}(x)] \pm z_{(1-\frac{\alpha}{2})} \times \sqrt{\hat{\text{Var}} [\text{logito} [\hat{\pi}(x)]]}.$$

Para se obter o intervalo de confiança para as probabilidades preditas pelo modelo, aplica-se a expressão (2.7) utilizando os limites inferior e superior do intervalo de confiança do logito. Intervalos de confiança para as razões de chances são obtidos de forma semelhante. Se a razão de chances é dada por (2.8) e tem-se intervalos de confiança para as estimativas dos parâmetros, o intervalo de confiança é dado por (AGRESTI, 2002):

$$\exp \left(\hat{\beta}_1 \pm z_{(1-\frac{\alpha}{2})} \times \sqrt{\hat{\text{Var}} [\hat{\beta}_1]} \right). \quad (2.9)$$

Por fim, vale lembrar que há ainda outras opções para a função de ligação $g(\mu)$. As duas mais conhecidas caracterizam os modelos conhecidos por probito, $\eta = \Phi^{-1}(\pi)$, em que Φ denota a distribuição normal padrão, e complemento log-log, em que $\eta = \ln [-\ln (1 - \pi)]$, mas a interpretação dos parâmetros β não pode mais ser feita em termos das razões de chances como no modelo logístico.

2.1.2.2 Variáveis politômicas

Quando a variável categorizada é politômica e os ensaios podem ser considerados independentes e idênticos, a distribuição a ser utilizada é a multinomial (AGRESTI, 2007). Seja o número de categorias denotado por K , cujas probabilidades de sucesso são denotadas por $\pi_1, \dots, \pi_K$ em que $\sum_{k=1}^K \pi_k = 1$, n ensaios e $y_1, \dots, y_K$ as quantidades de ensaios correspondentes

às K categorias de forma que $\sum_{k=1}^K y_k = n$. Assim, a distribuição multinomial tem função de probabilidade

$$P[y_1, y_2, \dots, y_K] = \left(\frac{n!}{y_1! y_2! \dots y_K!} \right) \pi_1^{y_1} \pi_2^{y_2} \dots \pi_K^{y_K}.$$

Para a distribuição multinomial a função de ligação canônica é

$$\eta_k = \ln \left(\frac{\pi_k}{\pi_K} \right), \quad k = 1, \dots, K-1.$$

A distribuição binomial é um caso particular quando $K = 2$. A distribuição multinomial é multivariada, mas a distribuição marginal de uma categoria em particular é binomial.

Os modelos para dados politômicos nominais são uma extensão do caso dicotômico apresentado na seção 2.1.2.1 que, com o uso da função de ligação logito, são chamados de modelos de logitos generalizados. Estes modelos comparam cada categoria da variável resposta com uma categoria fixada K (também chamada de base). Deste modo, têm-se $(K - 1)$ logitos, cada um com um vetor de parâmetros associados:

$$\text{logito}[\pi_k] = \ln \left(\frac{\pi_k}{\pi_K} \right) = \mathbf{x}' \boldsymbol{\beta}_k. \quad (2.10)$$

As probabilidades preditas pelo modelo de logitos generalizados são agora dadas por

$$\pi_k(\mathbf{x}) = \frac{\exp(\mathbf{x}' \boldsymbol{\beta}_k)}{1 + \sum_{h=1}^{K-1} \exp(\mathbf{x}' \boldsymbol{\beta}_h)}. \quad (2.11)$$

Como ilustração, considere uma variável com $K = 3$ categorias descrita por um modelo com apenas uma variável *dummy* (X), indicando presença/ausência de uma característica ($X = 1$ e $X = 0$, respectivamente). O modelo é expresso pelos logitos

$$\text{logito}[\pi_1] = \ln \left(\frac{\pi_1}{\pi_3} \right) = \beta_{01} + \beta_{11}X$$

e

$$\text{logito}[\pi_2] = \ln \left(\frac{\pi_2}{\pi_3} \right) = \beta_{02} + \beta_{12}X,$$

fixando-se a categoria $k = 3$ como base (referência). As probabilidades de ocorrência das categorias $k = 1$, $k = 2$ e $k = 3$ para uma observação quando $X = 1$ são dadas por

$$\hat{\pi}_1 = \frac{\exp(\hat{\beta}_{01} + \hat{\beta}_{11})}{1 + \exp(\hat{\beta}_{01} + \hat{\beta}_{11}) + \exp(\hat{\beta}_{02} + \hat{\beta}_{12})}, \quad (2.12)$$

$$\hat{\pi}_2 = \frac{\exp(\hat{\beta}_{02} + \hat{\beta}_{12})}{1 + \exp(\hat{\beta}_{01} + \hat{\beta}_{11}) + \exp(\hat{\beta}_{02} + \hat{\beta}_{12})} \quad (2.13)$$

e

$$\hat{\pi}_3 = 1 - (\hat{\pi}_1 + \hat{\pi}_2),$$

uma vez que $\text{logito}[\hat{\pi}_k] = \hat{\beta}_{0k} + \hat{\beta}_{1k}$, para $k = 1, 2$. A variância do k -ésimo logito é dada por

$$\hat{\text{Var}}[\text{logito}[\hat{\pi}_k]] = \hat{\text{Var}}[\hat{\beta}_{0k}] + \hat{\text{Var}}[\hat{\beta}_{1k}] + 2\hat{\text{cov}}[\hat{\beta}_{0k}, \hat{\beta}_{1k}]$$

e, portanto, um intervalo de confiança para o k -ésimo logito é dado por

$$\text{logito}[\hat{\pi}_k] \pm z_{(1-\frac{\alpha}{2})} \times \sqrt{\hat{\text{Var}}[\text{logito}[\hat{\pi}_k]]}. \quad (2.14)$$

Com isso, os intervalos de confiança para as probabilidades π_1 e π_2 são obtidos aplicando-se os limites dos intervalos calculados para os logitos em (2.14) nas expressões (2.12) e (2.13), respectivamente.

Quanto às razões de chances utilizadas na interpretação dos parâmetros, estas são similares ao caso dicotômico. Porém, como há $(K - 1)$ logitos, cada um terá suas respectivas interpretações uma vez que os parâmetros são diferentes entre os logitos. No mesmo exemplo acima, $e^{\beta_{11}}$ representa a razão de chances entre as categorias $k = 1$ e $k = 3$ equivalente à variável X , pois

$$\psi = \frac{\pi_1(X = 1)/\pi_3(X = 1)}{\pi_1(X = 0)/\pi_3(X = 0)} = \frac{\exp(\beta_{01} + \beta_{11})}{\exp(\beta_{01})} = e^{\beta_{11}}.$$

Isso significa que a chance uma observação com $X = 1$ ser classificada na categoria $k = 1$ comparada à categoria $k = 3$ é $(e^{\beta_{11}})$ vezes a chance de uma observação com $X = 0$; a interpretação do parâmetro β_{12} é análoga, mas comparando as categorias $k = 2$ e $k = 3$. Ademais, um intervalo de confiança para essa razão de chances é obtido da mesma maneira mostrada em (2.9). Enquanto os intervalos para probabilidades dependem de todos os logitos presentes no modelo, os intervalos para as razões de chances dependem somente do logito referente à comparação representada por cada logito.

2.2 Estudos longitudinais

Ao contrário dos estudos transversais, em que apenas uma medida da variável resposta é feita em cada indivíduo (unidade amostral), nos estudos longitudinais podem ser tomadas várias mensurações de um mesmo indivíduo (sendo ao longo do tempo o caso mais frequente). A amostragem no caso longitudinal pode ser feita tanto retrospectiva quanto prospectivamente, embora a última seja mais confiável e aconselhável para assegurar a qualidade dos dados coletados. Entre as vantagens de estudos assim, está a possibilidade de se estudarem efeitos entre e/ou intra-indivíduos ao longo do tempo, o que não é possível no caso transversal. Outro ponto de destaque é que, com o mesmo número de indivíduos, atingem-se os mesmos objetivos com mais observações e por isso diz-se que os estudos longitudinais são mais poderosos (DIGGLE *et al.*, 2002; ZEGER; LIANG, 1992). Ademais, deve-se atentar para a diferença entre estudos longitudinais e aqueles denominados séries temporais: nos primeiros, geralmente, há várias unidades amostrais sendo observadas em ocasiões de tempo enquanto que, nestes últimos há apenas uma unidade amostral observada (SANTANA, 2004).

Estudos longitudinais são, às vezes, confundidos com os chamados estudos com medidas repetidas. Um exemplo de experimento com medidas repetidas pode ser pensado quando mede-se a capacidade auditiva de pessoas utilizando para isso observações feitas nos dois ouvidos; espera-se que tais observações estejam relacionadas por serem feitas em um mesmo indivíduo. Em contrapartida, um exemplo de estudo longitudinal é quando se mede a pressão sanguínea de indivíduos em várias ocasiões diferentes: nesse caso, o tempo acaba estabelecendo uma “ordem” em que as observações foram feitas. (VERBEKE *et al.*, 2014).

A estrutura de um conjunto de dados longitudinais para uma variável resposta contínua é representada na Tabela 2.1.

Tabela 2.1 – Estrutura de um conjunto de dados longitudinais com resposta contínua

Unidade amostral	Variável resposta				Variáveis explicativas			
	1	2	...	T	1	2	...	T
1	Y_{11}	Y_{12}	...	Y_{1T}	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$	...	$\mathbf{x}_{1T}$
2	Y_{21}	Y_{22}	...	Y_{2T}	$\mathbf{x}_{21}$	$\mathbf{x}_{22}$	...	$\mathbf{x}_{2T}$
3	Y_{31}	Y_{32}	...	Y_{3T}	$\mathbf{x}_{31}$	$\mathbf{x}_{32}$	...	$\mathbf{x}_{3T}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Y_{n1}	Y_{n2}	...	Y_{nT}	$\mathbf{x}_{n1}$	$\mathbf{x}_{n2}$	...	$\mathbf{x}_{nT}$

Os modelos tradicionais de regressão, como os MLG (seção 2.1.1) e os modelos lineares, não são mais aplicáveis a esta situação pois as observações, em geral, não podem ser consideradas independentes, pois as várias mensurações tendem a estar correlacionadas. Se esses métodos tradicionais forem aplicados neste caso, muito provavelmente haja distorções nas estimativas dos parâmetros e seus erros padrões, consequentemente causando inferências errôneas (AGRANONIK, 2009).

Os objetivos gerais de uma análise longitudinal podem ser resumidos em (ZEGGER; LIANG, 1992; VERBEKE *et al.*, 2014):

1. Descrever o comportamento da variável resposta de acordo com as covariáveis, assim como nos modelos clássicos de regressão, e com o tempo;
2. A fim de obter um cenário mais realístico, incorporar a informação das correlações intra-indivíduos nesses modelos.

As metodologias para o estudo de dados longitudinais dividem-se em três categorias (ZEGGER; LIANG, 1992): modelos marginais, de efeitos aleatórios e de transição (que não são abordados neste texto). Nas próximas seções discutem-se os modelos marginais e de efeitos aleatórios, suas particularidades e como se dá a estimação dos parâmetros, e por fim é feita uma comparação das vantagens e desvantagens de cada abordagem.

2.2.1 Modelos marginais

Um modelo marginal é aquele em que a regressão da variável resposta é modelada separadamente da correlação intra-indivíduos. Na etapa da regressão, a esperança marginal (média da variável resposta de observações que têm em comum algum conjunto de valores das covariáveis) é modelada em função das covariáveis. A esperança marginal é o foco dos estudos transversais (DIGGLE *et al.*, 2002). O termo “marginal” é designado ainda porque todos os indivíduos têm o mesmo efeito estimado para as covariáveis.

Um dos artigos pioneiros a tratar de modelos marginais para dados longitudinais foi proposto por Liang e Zeger (1986), como uma extensão dos modelos lineares generalizados em que é possível flexibilizar a matriz de variâncias e covariâncias a fim de incorporar as correlações entre as diversas observações de um mesmo indivíduo. A esta abordagem eles deram o nome de **equações de estimação generalizadas** (EEG ou GEE na sigla em inglês - *generalized estimating equations*), que será detalhada a seguir.

2.2.1.1 Equações de estimação generalizadas (EEG)

A proposta das EEG trata-se de uma extensão dos MLG para a análise de dados longitudinais quando o principal interesse está na regressão; desta forma, os parâmetros referentes à estrutura de correlação são tratados como *nuisance* (de perturbação) e como um problema secundário. O principal desafio em estudos longitudinais está na dificuldade em especificar uma distribuição de probabilidade conjunta para todos os instantes de tempo quando os dados não têm distribuição normal (para o caso normal isso é feito por meio da distribuição normal multivariada). Nas EEG, esta distribuição conjunta não é especificada e são utilizadas equações de estimação que proveem estimativas consistentes dos parâmetros de regressão e suas variâncias sob condições regulares (LIANG; ZEGGER, 1986).

Inicialmente, sejam n indivíduos ($i = 1, \dots, n$), cada um avaliado em T_i ocasiões diferentes ($t = 1, \dots, T_i$). Sejam então o vetor de respostas ($T_i \times 1$) do i -ésimo indivíduo $Y_i = (y_{i1}, \dots, y_{iT_i})'$ e uma matriz ($T_i \times p$) de covariáveis do i -ésimo indivíduo $X_i = (x_{i1}, \dots, x_{iT_i})'$. A função de densidade marginal de y_{it} exposta em (2.3) agora é

$$f(y_{it}; \theta_{it}, \phi) = \exp \left\{ \phi^{-1} [y_{it}\theta_{it} - b(\theta_{it})] + c(y_{it}, \phi) \right\} \quad i = 1, \dots, n \quad t = 1, \dots, T_i. \quad (2.15)$$

A esperança e a variância são dadas, analogamente, por

$$E[Y_{it}] = b'(\theta_{it}) \text{ e } \text{Var}[Y_{it}] = \phi b''(\theta_{it}).$$

O primeiro passo consiste no ajuste de um modelo de regressão assumindo que as observações de um mesmo indivíduo são independentes entre si. Na literatura estatística isto é chamado de *naïve*, indicando uma certa “ingenuidade” (tradução livre) ao não se levar em consideração a correlação que tende a surgir entre observações de um mesmo indivíduo. Surge

assim um estimador de β denotado por $\hat{\beta}_I$. Sob tal suposição, a função escore da função de verossimilhança tem a forma

$$U_I(\beta) = \sum_{i=1}^n X_i' \Delta S_i = 0, \quad (2.16)$$

em que $\Delta_i = \text{diag}(\partial \theta_{it} / \partial \eta_{it})$ é uma matriz $(T_i \times T_i)$ e $S_i = Y_i - \mu_i$ é um vetor $(T_i \times 1)$ do i -ésimo indivíduo. A solução da equação (2.16) é o estimador $\hat{\beta}_I$. Este estimador é consistente quando o modelo de regressão é especificado corretamente, conforme demonstrado por Liang e Zeger (1986). Contudo, $\hat{\beta}_I$ não tem uma alta eficiência quando a correlação entre as observações intra-indivíduo é alta, sendo tal deficiência corrigida nos passos seguintes do processo de ajuste via EEG.

Sejam também as matrizes $A_i = \text{diag}\{b''(\theta_{it})\}$ e $R_i(\alpha)$, esta chamada matriz de “correlação de trabalho”, tendo dimensão $(T_i \times T_i)$ e satisfazendo os critérios de uma matriz de correlações. A matriz

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} \quad (2.17)$$

será igual a $\text{cov}[Y_i]$ quando $R_i(\alpha)$ for de fato a verdadeira matriz de correlação das observações. As EEG são definidas por

$$U(\beta, \alpha) = \sum_{i=1}^n D_i' V_i^{-1} S_i = 0, \quad (2.18)$$

em que $D_i = \partial \mu_i / \partial \beta$. Cada $U_i(\beta, \alpha)$ é similar à função especificada na abordagem de quase-verossimilhança de Wedderburn (1974) e McCullagh (1983), exceto pelo fato de agora os V_i serem função também de α , além dos β . A expressão (2.18) pode ser reescrita como função dos β substituindo-se α em (2.17) e (2.18) por $\hat{\alpha}(Y, \beta, \phi)$, um estimador $\sqrt{n}$ -consistente de α quando β e ϕ forem conhecidos. Para completar o processo, ϕ é substituído por $\hat{\phi}(Y, \beta)$, um estimador $\sqrt{n}$ -consistente quando β for conhecido. Posto isso, a expressão (2.18) fica dada por

$$\sum_{i=1}^n U_i \left[\beta, \hat{\alpha} \left\{ \beta, \hat{\phi}(\beta) \right\} \right] = 0,$$

cujas soluções são denotadas por $\hat{\beta}_G$.

O estimador de α varia conforme a escolha da matriz $R_i(\alpha)$, mas uma forma geral de uma estimativa consistente é dada por

$$\hat{R}_{uv} = \sum_{i=1}^n \frac{\hat{r}_{iu} \hat{r}_{iv}}{N - p},$$

em que $N = \sum T_i$ e $\hat{r}_{it}$ é o resíduo de Pearson definido por

$$\hat{r}_{it} = \frac{y_{it} - \hat{\mu}_{it}}{\sqrt{V(\hat{\mu}_{it})}}.$$

Em seu artigo, Liang e Zeger (1986) afirmam que o método das EEG é robusto mesmo com a especificação errônea da matriz $R_i(\alpha)$ - os resultados encontrados são similares para

diferentes matrizes. Algumas das opções de estruturas que podem ser utilizadas são mostradas a seguir (considerando, a título de exemplo, seis observações para cada indivíduo) (TWISK, 2003; LARA, 2007).

A primeira possibilidade é a **estrutura independente**. A matriz $R_i(\alpha)$ é a identidade, o que equivale a assumir que não há correlação entre as várias observações de um mesmo indivíduo:

$$R_i(\alpha) = I_6 \quad \rightarrow \quad \begin{array}{c|cccccc} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \\ \hline t_1 & 1 & 0 & 0 & 0 & 0 & 0 \\ t_2 & 0 & 1 & 0 & 0 & 0 & 0 \\ t_3 & 0 & 0 & 1 & 0 & 0 & 0 \\ t_4 & 0 & 0 & 0 & 1 & 0 & 0 \\ t_5 & 0 & 0 & 0 & 0 & 1 & 0 \\ t_6 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

A segunda opção é chamada de **permutável**, na qual todas as correlações entre os tempos são consideradas iguais:

$$R_i(\alpha) = \begin{cases} 1 & \text{se } t = t' \\ \rho & \text{se } t \neq t' \end{cases} \quad \rightarrow \quad \begin{array}{c|cccccc} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \\ \hline t_1 & 1 & \rho & \rho & \rho & \rho & \rho \\ t_2 & \rho & 1 & \rho & \rho & \rho & \rho \\ t_3 & \rho & \rho & 1 & \rho & \rho & \rho \\ t_4 & \rho & \rho & \rho & 1 & \rho & \rho \\ t_5 & \rho & \rho & \rho & \rho & 1 & \rho \\ t_6 & \rho & \rho & \rho & \rho & \rho & 1 \end{array}$$

A estrutura chamada de **m -dependente** considera que as correlações com defasagens maiores que m são nulas, e as correlações com mesma defasagem são iguais entre si. Por exemplo, uma estrutura 2-dependente tem a seguinte forma:

$$R_i(\alpha) = \begin{cases} 0 & \text{se } |t - t'| > 2 \\ \rho_{|t-t'|} & \text{se } |t - t'| \leq 2 \end{cases} \quad \rightarrow \quad \begin{array}{c|cccccc} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \\ \hline t_1 & 1 & \rho_1 & \rho_2 & 0 & 0 & 0 \\ t_2 & \rho_1 & 1 & \rho_1 & \rho_2 & 0 & 0 \\ t_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & 0 \\ t_4 & 0 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ t_5 & 0 & 0 & \rho_2 & \rho_1 & 1 & \rho_1 \\ t_6 & 0 & 0 & 0 & \rho_2 & \rho_1 & 1 \end{array}$$

Outra estrutura muito utilizada é a **autorregressiva** (AR-1). Nela, as correlações defasadas de t tempos são consideradas como ρ^t :

$$R_i(\alpha) = \begin{cases} 1 & \text{se } t = t' \\ \rho^{|t-t'|} & \text{se } t \neq t' \end{cases} \rightarrow$$

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	1	ρ^1	ρ^2	ρ^3	ρ^4	ρ^5
t_2	ρ^1	1	ρ^1	ρ^2	ρ^3	ρ^4
t_3	ρ^2	ρ^1	1	ρ^1	ρ^2	ρ^3
t_4	ρ^3	ρ^2	ρ^1	1	ρ^1	ρ^2
t_5	ρ^4	ρ^3	ρ^2	ρ^1	1	ρ^1
t_6	ρ^5	ρ^4	ρ^3	ρ^2	ρ^1	1

Por último, há a opção **não-estruturada**. Nesse caso todas as correlações são assumidas serem diferentes umas das outras, mas isso implica em um grande número de parâmetros a serem estimados e faz dessa estrutura uma opção pouco utilizada, somente em casos extremos:

$$R_i(\alpha) = \begin{cases} 1 & \text{se } t = t' \\ \rho_{tt'} & \text{se } t \neq t' \end{cases} \rightarrow$$

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	1	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{15}	ρ_{16}
t_2	ρ_{12}	1	ρ_{23}	ρ_{24}	ρ_{25}	ρ_{26}
t_3	ρ_{13}	ρ_{23}	1	ρ_{34}	ρ_{35}	ρ_{36}
t_4	ρ_{14}	ρ_{24}	ρ_{34}	1	ρ_{45}	ρ_{46}
t_5	ρ_{15}	ρ_{25}	ρ_{35}	ρ_{45}	1	ρ_{56}
t_6	ρ_{16}	ρ_{26}	ρ_{36}	ρ_{46}	ρ_{56}	1

Diferentes estruturas de correlação adotadas para a matriz $R_i(\alpha)$ afetam pouco as estimativas dos parâmetros β , apenas os respectivos erros padrões. Como sugere Twisk (2003), uma primeira forma de escolher tal estrutura é analisar a matriz de correlações observada entre os tempos (embora isso não faça tanto sentido para respostas categorizadas). Outro fator importante a ser levado em consideração é a parcimônia, ou seja, escolher a estrutura com menor número de parâmetros a serem estimados, desde que haja sentido prático em adotá-la. Para casos em que as respostas são categorizadas há a possibilidade de utilizar razões de chances para captar a associação entre as observações (VERBEKE *et al.*, 2014).

Um critério adicional que auxilia tanto na escolha da estrutura de correlação quanto na seleção de variáveis é o *quasi-likelihood information criterion* (QIC), proposto por Pan (2001). Trata-se de uma modificação do tradicional critério AIC (*Akaike's information criterion*) (AKAIKE, 1974), uma vez que este não pode ser utilizado no contexto de EEG por ser baseado na teoria de máxima verossimilhança; Pan (2001) utiliza a quase-verossimilhança no QIC. Assim como o AIC, valores baixos de QIC indicam que um modelo é adequado, porém dificilmente os *softwares* têm este critério implementado.

No *software* R (R Core Team, 2014) o ajuste de modelos marginais via EEG pode ser feito por meio dos pacotes *geepack* (HØJSGAARD; HALEKOH; YAN, 2006) e *gee* (CAREY; LUMLEY; RIPLEY, 2012), que ajustam modelos para respostas com distribuições pertencentes à família dos MLG e são alguns dos mais utilizados. No *software* SAS (SAS, 2011) existe o procedimento PROC GENMOD (SAS, 2008a) para este fim, quando o usuário informa a opção *repeated*.

2.2.2 Modelos de efeitos aleatórios

Os modelos de efeitos aleatórios são também chamados de modelos mistos pois consideram em sua estrutura, além de uma parte fixa (constantes ou termos associados a certos níveis de um fator, por exemplo), efeitos aleatórios para os quais se assumem distribuições de probabilidade (por isso a denominação **aleatório**); no entanto, a nomenclatura de “modelos mistos” é reservada àqueles modelos que tenham termos aleatórios além do erro (excluindo-se assim os modelos de regressão linear, por exemplo). A inclusão de efeitos aleatórios tem por objetivo, geralmente, captar e explicar dependências entre observações que possam surgir dependendo do tipo de estudo ou delineamento realizado. Em dados longitudinais, tal inclusão é utilizada para explicar a dependência entre as observações feitas em um mesmo indivíduo. Com a presença de efeitos aleatórios em um modelo, a matriz de covariâncias das variáveis respostas considera uma correlação representando as observações relacionadas entre si (PINHEIRO; BATES, 2000; SILVA, 2013).

A ideia central dos modelos mistos é que os parâmetros do modelo de regressão podem variar entre os indivíduos de acordo com uma distribuição de probabilidades, ou seja, há uma heterogeneidade considerável entre estes parâmetros. Assumindo-se que as observações de um indivíduo representam realizações de uma trajetória específica dele ao longo do tempo, e supondo que os indivíduos representam uma amostra aleatória da população, consequentemente os parâmetros específicos de um indivíduo são também aleatórios e por isso dá-se o nome de **efeitos aleatórios**. A dependência entre as observações de um indivíduo surge dos efeitos aleatórios em comum entre estas observações. Os modelos mistos são conhecidos também por outras denominações como modelos hierárquicos, multiníveis ou ainda de componentes de variância dependendo do contexto em que são aplicados (DIGGLE *et al.*, 2002; VERBEKE *et al.*, 2014).

Um dos primeiros estudos de modelos mistos para dados longitudinais foi feito por Laird e Ware (1982), para casos em que a variável resposta é assumida tendo distribuição normal (estes modelos são conhecidos também por modelos lineares mistos - MLM - por tratar-se de extensões dos modelos lineares clássicos). Em seu trabalho, os autores propõem uma formulação do modelo em dois estágios que podem ser resumidos como a seguir.

Assim como definido anteriormente, seja Y_i o vetor de T_i observações do i -ésimo indivíduo. Tem-se o modelo

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i,$$

em que X_i é uma matriz de delineamento conhecida ($T_i \times p$) relacionada ao vetor β de efeitos fixos com dimensão ($p \times 1$); Z_i é uma matriz de delineamento conhecida ($T_i \times q$) relacionada ao vetor b_i de efeitos aleatórios a nível de indivíduo, com dimensão ($q \times 1$). Tem-se ainda que $\varepsilon_i \sim N(0, \Sigma_i)$ e $b_i \sim N(0, D)$, sendo os vetores b_i e ε_i independentes entre si. Os elementos das matrizes Σ_i e D são denominados componentes de variância.

Destacam-se os seguintes resultados acerca deste modelo:

i) Condicionado aos efeitos aleatórios $\mathbf{b}_i$, ou seja, fixando-se o i -ésimo indivíduo, $\mathbf{Y}_i$ tem distribuição normal com os seguintes parâmetros:

$$\mathbf{Y}_i|\mathbf{b}_i \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Sigma}_i).$$

Exemplificando, pode-se fazer $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}$, o que equivale a assumir que as T_i observações do indivíduo i são independentes com variância σ^2 .

ii) A distribuição marginal do vetor $\mathbf{Y}_i$, $f_i(\mathbf{y}_i)$, é obtida fazendo-se

$$f_i(\mathbf{y}_i) = \int f_i(\mathbf{y}_i|\mathbf{b}_i) \cdot f(\mathbf{b}_i) d\mathbf{b}_i, \quad (2.19)$$

que pode ser demonstrado que se trata também de uma distribuição normal:

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \boldsymbol{\Sigma}_i).$$

Ou seja, o modelo misto (hierárquico) implica em um modelo marginal para o vetor $\mathbf{Y}_i$ (MOLENBERGHS; VERBEKE, 2005).

A escolha por quais efeitos aleatórios serão colocados no modelo deve ser baseada aliando-se a análise exploratória dos dados e as interpretações que serão feitas acerca dos parâmetros. Quando se coloca um efeito aleatório em uma variável do modelo assume-se que esta age de formas distintas sob os indivíduos, i.e., seu efeito não é o mesmo para todos. Pode-se dizer que os dois principais tipos de modelos com efeitos aleatórios para dados longitudinais são os modelos de:

1. interceptos aleatórios;
2. interceptos e inclinações aleatórios.

Nos primeiros, é atribuído um efeito aleatório ao intercepto do modelo, fazendo com que haja diferentes curvas estimadas para cada indivíduo (além de uma curva média). Por exemplo, para um modelo apenas com a variável tempo tem-se:

$$Y_{it} = (\beta_0 + b_{0i}) + \beta_1 t_{it} + \varepsilon_{it}. \quad (2.20)$$

No segundo caso são considerados efeitos aleatórios pertinentes em outras covariáveis do modelo, fazendo com que as curvas passem a ter trajetórias específicas. Por exemplo, se em (2.20) agora atribui-se um efeito aleatório na variável tempo tem-se

$$Y_{it} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{it} + \varepsilon_{it}. \quad (2.21)$$

Nota-se claramente que neste modelo a variável t_{ij} passa a ter diferentes efeitos sob cada um dos i indivíduos.

Os dois principais métodos de estimação dos parâmetros do modelo, dentro do contexto frequentista, são a máxima verossimilhança (MV) e da máxima verossimilhança restrita (conhecida por REML, na sigla em inglês). Entre as vantagens deste último sabe-se que os estimadores dos componentes de variância presentes nas matrizes $\mathbf{D}$ e Σ_i são não-viciados (quando os dados são balanceados). Em se tratando de métodos de seleção de modelos, podem ser utilizados testes de razão de verossimilhanças tanto para os efeitos fixos quando aleatórios, desde que os modelos sejam encaixados (VERBEKE; MOLENBERGHS, 2000).

Porém, em muitos casos, a variável resposta de interesse não tem distribuição normal, como descrito anteriormente. Neste contexto, devem ser considerados, portanto, os modelos lineares generalizados mistos (MLGM), uma combinação dos MLG (seção 2.1.1) e da inclusão de efeitos aleatórios no modelo (analogamente aos MLM). Seguindo a notação adotada até aqui, assume-se nos MLGM que a distribuição de Y_{it} condicional aos efeitos aleatórios $\mathbf{b}_i$ é dada por

$$f(y_{it}|\mathbf{b}_i; \theta_{it}, \phi) = \exp \left\{ \phi^{-1} [y_{it}\theta_{it} - b(\theta_{it})] + c(y_{it}, \phi) \right\} \quad i = 1, \dots, n \quad t = 1, \dots, T_i,$$

similarmente à expressão (2.15). Também é usual (mas não restrita) a suposição de uma distribuição normal para os efeitos aleatórios, isto é, $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$. A média μ_{ij} continua sendo descrita pela função de ligação e um preditor linear, contendo efeitos fixos mas também efeitos aleatórios ($\mathbf{b}_i$), ou seja,

$$g(\mu_{it}) = g(E[Y_{it}|\mathbf{b}_i]) = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{b}_i.$$

A suposição de uma distribuição normal aos efeitos aleatórios ($f(\mathbf{b}_i) \sim N(\mathbf{0}, \mathbf{D})$), como descrito aqui, é a situação mais comum na prática e mostra-se adequada na maior parte dos casos. Nos MLM, Verbeke e Molenberghs (2000) concluem que a escolha da distribuição dos efeitos aleatórios não tem grande influência na estimação dos parâmetros da parte fixa do modelo; Zeger e Liang (1992) descrevem que para dados discretos longitudinais com poucas observações intra unidade amostral é difícil avaliar com segurança a melhor distribuição, e também sobre a dificuldade em fazer tal escolha.

A distribuição marginal de $\mathbf{Y}_i$ é dada pela expressão (2.19). Como a distribuição de $\mathbf{Y}_i$ condicional aos $\mathbf{b}_i$ é

$$f_i(\mathbf{y}_i|\mathbf{b}_i) = \prod_{t=1}^{T_i} f_{it}(\mathbf{y}_{it}|\mathbf{b}_i), \quad (2.22)$$

a expressão (2.19) passa a ser

$$f_i(\mathbf{y}_i) = \int \prod_{t=1}^{T_i} f_{it}(\mathbf{y}_{it}|\mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i.$$

Portanto, a verossimilhança para $\boldsymbol{\beta}$, $\mathbf{D}$ e ϕ é dada por:

$$L(\boldsymbol{\beta}, \mathbf{D}, \phi) = \prod_{i=1}^n f_i(\mathbf{y}_i) = \prod_{i=1}^n \int \prod_{t=1}^{T_i} f_{it}(\mathbf{y}_{it}|\mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i. \quad (2.23)$$

A estimação dos parâmetros do MLGM pode ser feita pelo método da máxima verossimilhança, o que implica maximizar a expressão (2.23). Contudo, a presença das n integrais sob os efeitos aleatórios torna esta tarefa mais complicada em relação aos MLM e de alguns outros casos específicos que têm solução analítica (DIGGLE *et al.*, 2002).

Quanto aos efeitos aleatórios, é necessário obter suas predições (estimativas) para a obtenção dos valores preditos de cada indivíduo, bem como para o estudo da variabilidade entre indivíduos caso necessário. Assim como nos MLM, essas predições são baseadas na sua distribuição *a posteriori*, ou seja,

$$f_i(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\beta}, \mathbf{D}, \phi) = \frac{f_i(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\beta}, \phi)f(\mathbf{b}_i|\mathbf{D})}{\int f_i(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\beta}, \phi)f(\mathbf{b}_i|\mathbf{D})d\mathbf{b}_i}.$$

Porém, ao contrário dos MLM, esta distribuição não é uma normal e nos MLGM a moda é utilizada como um estimador para os $\mathbf{b}_i$, ao invés da média. Em outras palavras, o estimador $\hat{\mathbf{b}}_i$ é o valor de $\mathbf{b}_i$ que maximiza $f_i(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\beta}, \mathbf{D}, \phi)$, utilizando-se as estimativas dos parâmetros desconhecidos. Esse processo é denominado na literatura por “estimativas empíricas de Bayes” (EB) (MOLENBERGHS; VERBEKE, 2005).

As abordagens para a maximização de (2.23) consistem, resumidamente, em (BRESLOW; CLAYTON, 1993; MOLENBERGHS; VERBEKE, 2005):

1. **Aproximações do integrando:** o objetivo é obter aproximações da função a ser integrada, geralmente utilizando o método de Laplace ou modificações dele.
2. **Aproximações dos dados:** os dados são decompostos em termos da média e um termo referente ao erro, utilizando uma expansão em séries de Taylor que é uma função não-linear do preditor linear. Ou seja,

$$Y_{it} = \mu_{it} + \varepsilon_{it} = h(\mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{z}'_{it}\mathbf{b}_i) + \varepsilon_{it},$$

em que $h(\cdot)$ é a inversa da função de ligação e os erros têm uma distribuição com variância igual a $\text{Var}[Y_{it}|\mathbf{b}_i] = \phi V(\mu_{it})$, sendo $V(\cdot)$ a função de variância. Os métodos desenvolvidos para esta ótica variam sobre a forma com que a aproximação da média μ_{it} é feita, mas sempre baseando-se em expansões em séries de Taylor.

3. **Aproximações da integral:** a última proposta consiste em aproximar o resultado da integral, desde que ela seja da forma

$$\int f(z)\phi(z)dz, \tag{2.24}$$

sendo $f(\cdot)$ uma função conhecida e $\phi(\cdot)$ a densidade de uma distribuição normal padrão. Logo, primeiramente deve-se padronizar os efeitos aleatórios $\mathbf{b}_i$ apresentados anteriormente

de modo que satisfaçam essa condição fazendo-se $\delta_i = \mathbf{D}^{-1/2} \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{I})$. O preditor linear passa a ser então $\eta_{it} = \mathbf{x}'_{it} \boldsymbol{\beta} + \mathbf{z}'_{it} \mathbf{D}^{1/2} \delta_i$. Disso,

$$f_i(\mathbf{y}_i) = \int \prod_{t=1}^{T_i} f_{it}(\mathbf{y}_{it} | \mathbf{b}_i) f(\mathbf{b}_i) d\mathbf{b}_i = \int \prod_{t=1}^{T_i} f_{it}(\mathbf{y}_{it} | \delta_i) f(\delta_i) d\delta_i,$$

que tem a forma apropriada (2.24).

Na quadratura Gaussiana a integral (2.24) é aproximada por

$$\int f(z) \phi(z) dz \approx \sum_{q=1}^Q w_q f(z_q),$$

sendo Q a ordem da aproximação, mais precisa quanto maior for o valor de Q . Os z_q são chamados de nós ou pontos de quadratura e são soluções do polinômio de Hermite de ordem Q , encontrados em tabelas específicas juntamente com os w_q , denominados pesos.

A principal desvantagem da quadratura Gaussiana é que, dependendo do suporte de $f(z)$ os nós podem não pertencer à região de interesse, sendo conveniente reescalar esses nós de modo que mais nós sejam úteis. Esta modificação é chamada de quadratura Gaussiana adaptativa, em que os nós são tomados como se $f(z)\phi(z)$ fosse uma distribuição normal de média igual à moda de $\ln[f(z)\phi(z)]$ e variância

$$\left[-\frac{\partial^2}{\partial z^2} \ln[f(z)\phi(z)] \Big|_{z=\hat{z}} \right]^{-1}.$$

Em geral, a quadratura Gaussiana adaptativa utiliza menos nós do que a convencional, mas é computacionalmente mais intensiva.

Como o modelo é ajustado utilizando-se a máxima verossimilhança, as inferências acerca dos parâmetros são feitas baseando-se no fato que os estimadores têm, assintoticamente, distribuição normal:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbf{K}^{-1}),$$

sendo $\mathbf{K}$ a matriz de informação de Fisher. Isto possibilita a realização de testes como o teste de Wald, da razão de verossimilhanças e a construção de intervalos de confiança para os parâmetros (MOLENBERGHS; VERBEKE, 2005).

2.2.3 Diferenças entre modelos marginais e modelos de efeitos aleatórios

Até o presente momento foram apresentados os principais aspectos acerca dos modelos marginais e mistos. Esta seção destaca as principais diferenças entre as duas abordagens, evidenciando fatores que podem interferir diretamente na escolha de qual abordagem é mais indicada para cada situação prática.

É importante compreender, inicialmente, que ambas as abordagens são extensões dos MLG desenvolvidas para tratar do caso de medidas correlacionadas (como os dados longitudinais),

porém com ênfases e interpretações distintas. A forma como os modelos são construídos, ou seja, as suposições teóricas feitas acerca deles e que conseqüentemente interferem na interpretação dos parâmetros de cada um. Além disso, dependendo do conjunto de dados a ser analisado, uma ou outra abordagem mostra-se mais adequada.

O primeiro ponto de atenção diz respeito à estrutura dos dados em estudo. Quando os dados a serem analisados apresentam altos níveis de desbalanceamento, valores faltantes, diferentes números de observações por indivíduo ou ainda espaçamentos desiguais entre os tempos os modelos de efeitos aleatórios têm maior flexibilidade de ajuste a eles. Os modelos marginais não são tão versáteis nesses quesitos, pois os algoritmos de estimação podem apresentar problemas (LOCASCIO; ATRI, 2011).

Como discutido por Diggle *et al.* (2002) e Verbeke *et al.* (2014), a ideia fundamental de um modelo de efeitos aleatórios é que os parâmetros da regressão de cada indivíduo têm uma grande variabilidade entre si e comportam-se segundo alguma distribuição de probabilidades. Embora nos modelos marginais essa noção de “variabilidade” entre os indivíduos também esteja presente (ajusta-se uma curva média em torno da qual estão distribuídos os indivíduos) isso não é explicitamente levado em consideração como no modelo misto, em que se assume uma distribuição de probabilidades para os efeitos aleatórios. Ainda, por serem abordagens com fundamentos diferentes, os parâmetros dos modelos ajustados não podem ser comparados entre si quando se ajustam modelos com o mesmo preditor linear; aliás, em muitas situações, há grandes diferenças entre as estimativas dos parâmetros, evidenciando que não são comparáveis. Os modelos marginais e de efeitos aleatórios são conhecidos também por modelos PA (*population average*) e SS (*subject specific*), respectivamente (ZEGGER; LIANG; ALBERT, 1988; NEUHAUS; KALBFLEISCH; HAUCK, 1991).

Ademais, uma diferença importante entre tais modelos é quanto às interpretações dos parâmetros, como discutido por Molenberghs e Verbeke (2005), Parzen *et al.* (2011), Neuhaus, Kalbfleisch e Hauck (1991). Para ilustrar, considere um MLGM de interceptos aleatórios com variável resposta tendo distribuição Bernoulli, isto é, $Y_{it}|b_i \sim \text{Bernoulli}(\pi_{it})$:

$$\ln \left(\frac{\pi_{it}}{1 - \pi_{it}} \right) = (\beta_0 + b_i) + \beta_1 t_{it}. \quad (2.25)$$

As probabilidades de cada indivíduo em função dos tempos são obtidas tomando-se a esperança condicional aos efeitos aleatórios, ou seja, $E[Y_{it}|b_i] = \pi_{it}|b_i$:

$$\pi_{it}|b_i = \frac{\exp [(\beta_0 + b_i) + \beta_1 t_{it}]}{1 + \exp (\beta_0 + b_i + \beta_1 t_{it})}, \quad (2.26)$$

que são representadas pelas linhas mais finas da Figura 2.1. Porém, geralmente há o interesse em estudar o comportamento médio dos indivíduos (nesse exemplo, obter a probabilidade média de ocorrência do evento). Logo, esse efeito “marginal”, representado pela linha mais grossa da

Figura 2.1, é obtido fazendo-se

$$E[Y_{it}] = E[E[Y_{it}|b_i]] = E\left[\frac{\exp(\beta_0 + b_i + \beta_1 t_{it})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{it})}\right],$$

o que implica na resolução da integral

$$E[Y_{it}] = \int \frac{\exp(\beta_0 + b_i + \beta_1 t_{it})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{it})} f(b_i) db_i,$$

que é resolvida por métodos numéricos ou atribuindo valores simulados para os efeitos aleatórios. É importante ressaltar ainda que a este efeito marginal tem um significado distinto de obter o comportamento de um indivíduo com efeito aleatório nulo ($b_i = 0$), pois

$$E\left[\frac{\exp(\beta_0 + b_i + \beta_1 t_{it})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{it})}\right] \neq \frac{\exp(\beta_0 + \beta_1 t_{it})}{1 + \exp(\beta_0 + \beta_1 t_{it})} \quad (2.27)$$

uma vez que, em geral, $E[g(Y)] \neq g[E(Y)]$, em que $g(\cdot)$ é uma função qualquer. Ainda, na maioria dos casos de MLGM, o comportamento marginal não tem a mesma forma (distribuição) que o comportamento condicional aos efeitos aleatórios. No exemplo dado, isso significa que a média geral de todos os indivíduos não tem uma expressão analítica como em (2.26), dificultando a interpretação dos parâmetros.

Figura 2.1 – Representação de um modelo com interceptos aleatórios

Fonte: Verbeke, Molenberghs e Rizopoulos (2010)

Portanto, enquanto os modelos marginais (como aqueles ajustados via EEG) tem como resultado imediato o comportamento médio marginal para todos os indivíduos, nota-se que nos modelos mistos a interpretação deve ser condicionada aos efeitos aleatórios (tantos quanto forem colocados no modelo). Assim, a obtenção deste efeito médio pode tornar-se uma tarefa complicada, uma vez que nem sempre existe uma expressão fechada para tal e a complexidade aumenta conforme são adicionados efeitos aleatórios no preditor linear. No entanto, como mostra

a Figura 2.1, o efeito médio tem o mesmo “sentido” que os efeitos individuais, isto é, sabe-se que a curva “média” tem a mesma forma mas não se sabe a magnitude deste efeito. Para ilustrar a comparação entre as interpretações dos parâmetros apresenta-se a Figura 2.2, adaptada dos trabalhos de Verbeke, Molenberghs e Rizopoulos (2010) e Verbeke e Molenberghs (2014) em que se descreve uma situação experimental na qual os indivíduos são submetidos a dois tratamentos ao longo do tempo, para a qual foram ajustados um modelo via EEG e um MLGM de interceptos aleatórios. Observam-se na Figura 2.2 pequenas diferenças entre as curvas médias obtidas nos ajustes via EEG e MLGM, mas uma grande diferença quando se faz o efeito aleatório igual a zero, evidenciando o que foi discutido sobre a expressão (2.27).

Figura 2.2 – Comparação de um modelo marginal e de um modelo misto: (a) efeito marginal obtido em um MLGM, (b) efeito marginal obtido por um ajuste via EEG e (c) efeito observado para um indivíduo com efeito aleatório nulo

Fonte: Verbeke, Molenberghs e Rizopoulos (2010), Verbeke e Molenberghs (2014)

Logicamente, a interpretação dos parâmetros dos efeitos fixos também sofre alterações nos MLGM. Como já mencionado, conhece-se apenas a expressão da esperança condicional aos efeitos aleatórios (i.e., dos indivíduos). Isso implica que os parâmetros dos efeitos fixos devem ser interpretados a nível de indivíduo e não mais diretamente como um efeito médio como nos modelos marginais. A título de exemplo, estende-se o modelo (2.25) para

$$\ln \left(\frac{\pi_{it}}{1 - \pi_{it}} \right) = \beta_0 + b_i + \beta_1 t_{it} + \beta_2 X_i, \quad (2.28)$$

ou seja, inclui-se uma variável X_i que é invariante ao longo do tempo. Para interpretar o parâmetro β_1 , a razão de chances correspondente ao acréscimo de uma unidade de tempo é, fixando-se o valor de X_i ,

$$\psi = \frac{\exp(\beta_0 + b_i + \beta_1(t+1) + \beta_2 X_i)}{\exp(\beta_0 + b_i + \beta_1 t + \beta_2 X_i)},$$

que só pode ser simplificada para

$$\psi = \exp(\beta_1)$$

desde que seja fixado o valor de b_i , ou seja, fixado o i -ésimo indivíduo. Isso não significa afirmar que cada indivíduo (i) tem um efeito diferente referente ao parâmetro β_1 ; como pode-se notar na Figura 2.1 todas as linhas dos indivíduos têm a mesma inclinação (mesmo efeito de β_1).

Entretanto, o efeito médio na população não é o mesmo que para cada indivíduo (como ocorre nos modelos marginais). Por isso, diz-se que a interpretação de um MLGM é feita a nível de indivíduo.

Por último, se é necessário “fixar” um indivíduo, vale observar que tal interpretação passa a fazer sentido somente para os efeitos fixos referentes a variáveis que variam no tempo para um mesmo indivíduo (como o próprio tempo, exemplificado acima). Outros exemplos de variáveis assim são idade, peso, altura (admitindo que o indivíduo está em fase de crescimento). Em contrapartida, variáveis como gênero e nacionalidade são invariantes no tempo. Em último caso, a interpretação do parâmetro de uma variável assim ainda pode ser feita se forem comparados dois indivíduos i e i' , em que a variável em questão muda de valor entre eles. Por exemplo, no modelo (2.28), se X_i é uma *dummy* com valor 1 para o i -ésimo indivíduo e 0 para o i' -ésimo indivíduo, tem-se a razão de chances

$$\psi = \frac{\exp(\beta_0 + b_i + \beta_1 t + \beta_2)}{\exp(\beta_0 + b_{i'} + \beta_1 t)},$$

que, se forem fixados os efeitos aleatórios b_i e $b_{i'}$, passa a ser

$$\psi = \exp(\beta_2 + (b_i - b_{i'})).$$

Logo, há um confundimento entre os efeitos e a única forma de isolar β_2 é assumir que $b_i = b_{i'}$, o que não é diretamente observável nos dados e torna difícil a interpretação deste efeito referente a uma variável invariante no tempo para um mesmo indivíduo.

2.3 Extensões para respostas politômicas nominais longitudinais

Esta seção tem por objetivo apresentar os principais modelos desenvolvidos para dados politômicos nominais com estrutura longitudinal, utilizando a teoria exposta nas seções 2.1.2.2, 2.2.1 e 2.2.2.

2.3.1 Extensão das EEG para o caso politômico

2.3.1.1 Metodologia

Na década de 1990, Lipsitz, Kim e Zhao (1994) desenvolveram uma extensão das EEG descritas na seção 2.2.1.1 para respostas politômicas nominais ou ordinais. Como trata-se de uma extensão, este método reduz-se às EEG de Liang e Zeger (1986) quando a resposta é dicotômica.

Estendendo a notação adotada na seção 2.2.1.1, considere agora que a resposta Y é categorizada e composta por K níveis ($k = 1, \dots, K$). Sejam então K variáveis indicadoras Y_{ikt} em que $Y_{ikt} = 1$ se o indivíduo i tem a resposta k no instante t e $Y_{ikt} = 0$ caso contrário, dando origem a um vetor $\mathbf{Y}_{it} = [Y_{i1t}, \dots, Y_{i,K-1,t}]'$ de dimensão $(K - 1) \times 1$ para cada indivíduo i em cada tempo t . Nota-se que quando $K = 2$ o vetor $\mathbf{Y}_{it}$ reduz-se a um escalar, e que o elemento Y_{iKt} é omitido em $\mathbf{Y}_{it}$ pois $Y_{i+} = 1$ (o símbolo $+$ indica o total do índice k).

A distribuição marginal de $\mathbf{Y}_{it}$ é multinomial, isto é,

$$f(\mathbf{y}_{it}|\mathbf{x}_{it}, \boldsymbol{\beta}) \propto \prod_{k=1}^K \pi_{ikt}^{y_{ikt}}, \quad (2.29)$$

em que $\pi_{ikt} = \pi_{ikt}(\boldsymbol{\beta}) = E[Y_{ikt}|\mathbf{x}_{it}, \boldsymbol{\beta}] = P[Y_{ikt} = 1|\mathbf{x}_{it}, \boldsymbol{\beta}]$ é a probabilidade de ocorrência da resposta k no instante t . Por se tratar de um modelo marginal, o objetivo são as inferências acerca dos parâmetros $\boldsymbol{\beta}$ associados ao vetor $E[\mathbf{Y}_{it}|\mathbf{X}_i] = \boldsymbol{\pi}_{it}(\boldsymbol{\beta}) = [\pi_{i1t}, \dots, \pi_{i,K-1,t}]'$ de dimensão $(K-1) \times 1$, sendo que tais vetores podem ainda ser agrupados formando outro vetor $E[\mathbf{Y}_i|\mathbf{X}_i] = \boldsymbol{\pi}_i = [\boldsymbol{\pi}'_{i1}, \dots, \boldsymbol{\pi}'_{iT_i}]'$ de dimensão $T_i(K-1) \times 1$, em que $\mathbf{Y}_i = [\mathbf{Y}'_{i1}, \dots, \mathbf{Y}'_{iT_i}]'$, representando todas as observações do i -ésimo indivíduo (LIPSITZ; KIM; ZHAO, 1994).

Os parâmetros $\boldsymbol{\beta}$ são estimados por meio das EEG definidas na expressão 2.18. Neste caso, como a distribuição de $\mathbf{Y}_{it}$ é multinomial, a matriz $\mathbf{V}_i$ é formada por blocos $(K-1) \times (K-1)$ dados por

$$\mathbf{V}_{it} = \mathbf{V}_{it}(\boldsymbol{\beta}) = \text{Var}[\mathbf{Y}_{it}] = \text{diag}\{\boldsymbol{\pi}_{it}\} - \boldsymbol{\pi}_{it}\boldsymbol{\pi}_{it}',$$

em que $\text{diag}\{\boldsymbol{\pi}_{it}\}$ é uma matriz diagonal com os elementos de $\boldsymbol{\pi}_{it}$ na diagonal.

Da expressão 2.17, sabe-se que $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$. Para o caso multinomial, a matriz $\mathbf{R}_i(\boldsymbol{\alpha})$ passa a ser dada por

$$\mathbf{R}_i(\boldsymbol{\alpha}) = \begin{bmatrix} \mathbf{A}_{i1}^{-\frac{1}{2}} \mathbf{V}_{i1} \mathbf{A}_{i1}^{-\frac{1}{2}} & \boldsymbol{\rho}_{i12} & \cdots & \boldsymbol{\rho}_{i1T} \\ \boldsymbol{\rho}_{i21} & \mathbf{A}_{i2}^{-\frac{1}{2}} \mathbf{V}_{i2} \mathbf{A}_{i2}^{-\frac{1}{2}} & \cdots & \boldsymbol{\rho}_{i2T} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\rho}_{iT1} & \boldsymbol{\rho}_{iT2} & \cdots & \mathbf{A}_{iT_i}^{-\frac{1}{2}} \mathbf{V}_{iT_i} \mathbf{A}_{iT_i}^{-\frac{1}{2}} \end{bmatrix},$$

em que cada bloco $\mathbf{A}_{it}^{-\frac{1}{2}} \mathbf{V}_{it} \mathbf{A}_{it}^{-\frac{1}{2}}$ tem dimensão $(K-1) \times (K-1)$, tendo os elementos da diagonal iguais a 1 pois $\text{cor}(Y_{ikt}, Y_{ikt}) = 1$. Os elementos fora desta diagonal são dados por

$$\text{cor}(Y_{ijt}, Y_{ikt}) = \frac{\text{cov}(Y_{ijt}, Y_{ikt})}{[\pi_{ijt}(1 - \pi_{ijt})\pi_{ikt}(1 - \pi_{ikt})]^{-\frac{1}{2}}} = \frac{-\pi_{ijt}\pi_{ikt}}{[\pi_{ijt}(1 - \pi_{ijt})\pi_{ikt}(1 - \pi_{ikt})]^{-\frac{1}{2}}}.$$

Os demais elementos de $\mathbf{R}_i(\boldsymbol{\alpha})$, as matrizes $\boldsymbol{\rho}_{ist}$ (em que s e t são dois tempos distintos), têm a forma

$$\boldsymbol{\rho}_{ist}(\boldsymbol{\alpha}) = \mathbf{A}_{is}^{-\frac{1}{2}} E[(\mathbf{Y}_{is} - \boldsymbol{\pi}_{is})(\mathbf{Y}_{it} - \boldsymbol{\pi}_{it})'] \mathbf{A}_{it}^{-\frac{1}{2}}$$

e são os únicos elementos desconhecidos, uma vez que os blocos $\mathbf{A}_{it}^{-\frac{1}{2}} \mathbf{V}_{it} \mathbf{A}_{it}^{-\frac{1}{2}}$ são especificados completamente por $\boldsymbol{\pi}_i(\boldsymbol{\beta})$.

Portanto, para estimar $\mathbf{V}_i$ é necessário estimar $\boldsymbol{\rho}_{ist}(\boldsymbol{\alpha})$, fazendo uma parametrização semelhante à feita por Liang e Zeger (1986) para respostas binárias. Alguns exemplos são as estruturas expostas na seção 2.2.1.1, como a permutável ($\boldsymbol{\rho}_{ist} = \boldsymbol{\rho}$), m -dependente, não-estruturada, entre outras.

2.3.1.2 Recursos computacionais

Esta metodologia foi implementada computacionalmente por Williamson, Lipsitz e Kim (1999) por meio da macro do *software* SAS (SAS, 2011) denominada GEECAT, que ajusta modelos para dados nominais ou ordinais. Esta macro está implementada atualmente no procedimento PROC GENMOD (SAS, 2008a), mas apenas para o caso ordinal. Portanto, para a análise de dados nominais faz-se necessário o uso do código originalmente desenvolvido pelos autores. Contudo, conforme discutido também por Touloumis (2015) esta macro não está disponível publicamente, sendo necessário o contato com seus desenvolvedores para o acesso ao código.

No *software* R (R Core Team, 2014) não há, até o momento, pacotes desenvolvidos para esta metodologia. Entretanto há recursos disponíveis para uma metodologia semelhante que será detalhada na próxima seção.

2.3.2 EEG utilizando uma parametrização de razões de chances locais

2.3.2.1 Metodologia

No ajuste de modelos via EEG para dados politômicos (nominais ou ordinais) uma das principais dificuldades é descrever a estrutura de associação entre as observações de uma forma que faça sentido para variáveis categorizadas. Além disso, outra dificuldade é encontrar uma metodologia que se adeque a ambos os casos. Algumas parametrizações foram desenvolvidas para isso, como a extensão das EEG para o caso politômico apresentada anteriormente (seção 2.3.1) e outra utilizando razões de chances globais. Touloumis, Agresti e Kateri (2013) propuseram uma alternativa baseada em razões de chances locais. O objetivo dos autores foi estabelecer uma metodologia útil para respostas nominais e ordinais (ponto deficiente da metodologia de razões de chances globais, válida somente para o caso ordinal). As razões de chances locais têm grande aplicabilidade no estudo de tabelas de contingência, mas ainda não haviam sido utilizadas no contexto de EEG (TOULOUMIS; AGRESTI; KATERI, 2013).

A estrutura de associação entre as observações é definida a partir dos modelos de associação de Goodman (1985), mais especificamente os modelos RC, que se utilizam das razões de chances locais. O conjunto destas razões de chances formará o vetor α que, assim como definido no caso usual, tem por objetivo descrever o relacionamento entre as observações e também é tratado como *nuisance*.

O primeiro passo desta abordagem consiste em formar pares entre os tempos (t, t') : $(1, 2), (1, 3), \dots, (1, T), (2, 3), \dots, (2, T), \dots, (T - 1, T)$. Estes pares podem ser codificados, respectivamente, fazendo-se $c = 1, 2, \dots, T - 1, T, \dots, 2T - 3, \dots, L$. Dois instantes de tempo (t, t') podem ser cruzados em forma de tabelas de contingência, contando-se o número de observações presentes em cada categoria. Seja, por exemplo, um experimento com resposta politômica de 3 níveis (A, B e C) e 3 instantes de tempo. Cruzando-se os tempos t e t' (ignorando-se as

covariáveis) tem-se a Tabela 2.2, em que $F_{kk'}$ representa o número de observações da casela (k, k') .

Tabela 2.2 – Exemplo de tabela de contingência formada no cruzamento de dois instantes de tempo

		t'		
		A	B	C
t	A	F_{11}	F_{12}	F_{13}
	B	F_{21}	F_{22}	F_{23}
	C	F_{31}	F_{32}	F_{33}

Como há L tabelas como esta deve-se adicionar dois índices (t e t') a $F_{kk'}$ para indicar qual cruzamento está sendo considerado, ou seja, $F_{tk't'k'}$. Neste exemplo, como há 3 instantes de tempo, há $L = 3$ tabelas de contingência como a mostrada acima. Sejam agora as razões de chances locais

$$\psi_{tk't'k'} = \frac{F_{tk't'k'} \times F_{t,k+1,t',k'+1}}{F_{t,k+1,t',k'} \times F_{t,k,t',k'+1}},$$

em que $k, k' = 1, \dots, K - 1$ e $t < t' = 1, \dots, T$, que comparam as categorias duas a duas fornecendo uma medida de associação entre elas. E, por fim, seja α um vetor de $(L \times (K - 1)^2)$ razões de chances locais:

$$\alpha = (\psi_{1121}, \dots, \psi_{112(K-1)}, \dots, \psi_{(T-1)1T1}, \dots, \psi_{(T-1)(K-1)T(K-1)})'.$$

Porém, os referidos autores preferem trabalhar com uma outra reparametrização, que corresponde a uma generalização do modelo RC. As razões de chances locais, segundo este modelo, satisfazem

$$\ln(\psi_{tk't'k'}) = \gamma_{tt'} \left(\omega_{tk}^{tt'} - \omega_{t(k+1)}^{tt'} \right) \left(\omega_{t'k'}^{tt'} - \omega_{t'(k'+1)}^{tt'} \right), \quad (2.30)$$

em que os ω e γ são chamados de parâmetros escore e intrínsecos, respectivamente. No caso nominal, como não há uma ordenação entre as categorias da variável resposta, esses parâmetros devem ser estimados e algumas restrições devem ser feitas. Como sugerido por Touloumis, Agresti e Kateri (2013), tais restrições são

$$\sum_{k=1}^K \omega_k = 0 \quad \text{e} \quad \sum_{k=1}^K \omega_k^2 = 1.$$

A estimação dos parâmetros é feita em dois estágios. No primeiro estágio, para a estimação de α , as L tabelas de contingência são tratadas como independentes (embora não sejam), mas os autores defendem que isso não traz maiores problemas à consistência dos estimadores. Assume-se distribuições Poisson independentes para as contagens das tabelas de contingência formadas, e maximiza-se a verossimilhança da Poisson para estimá-las, adicionando-se uma pequena constante positiva em caso de muitas frequências nulas. A maximização da função de verossimilhança assegura a existência de $\hat{\alpha}$. No segundo estágio são solucionadas as equações escore $U(\beta, \hat{\alpha})$

similares àquelas especificadas em (2.18) para estimar os parâmetros β . Nota-se ainda que as razões de chances em (2.30) não dependem do modelo marginal especificado (parâmetros β).

As razões de chances locais podem ainda ser dispostas em uma matriz simétrica bloco-diagonal de dimensões $T(K-1) \times T(K-1)$ representada por

$$\begin{bmatrix} \mathbf{0} & \Psi_{12} & \cdots & \Psi_{1T} \\ \Psi_{21} & \mathbf{0} & \cdots & \Psi_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{T1} & \Psi_{T2} & \cdots & \mathbf{0} \end{bmatrix}.$$

Cada bloco tem dimensão $(K-1) \times (K-1)$ e cada elemento (k, k') dos blocos $\Psi_{tt'}$ é a estimativa de uma razão de chances local $\psi_{tk't'k'}$. Os blocos na diagonal são compostos de zeros pois refletem o fato que quando $t = t'$ não há razões de chances a serem estimadas (TOULOU MIS, 2015).

As duas estruturas para as razões de chances locais sugeridas para o caso nominal (outras estruturas propostas são adequadas apenas para o caso ordinal) são (TOULOU MIS; AGRESTI; KATERI, 2013; TOULOU MIS, 2015):

i) Estrutura permutável: esta estrutura não assume nenhuma dependência temporal e implica no fato de as razões de chances serem iguais para quaisquer pares de tempos (t, t') . Neste caso, a expressão (2.30) é simplificada para $\ln(\psi_{tk't'k'}) = \gamma (\omega_k - \omega_{k+1}) (\omega_{k'} - \omega_{k'+1})$. Recomenda-se o uso desta estrutura para tamanhos de amostra pequenos, devido a problemas de convergência.

ii) Estrutura RC: é uma extensão da opção anterior, mas que possibilita diferentes razões de chances entre os pares de tempos. Neste caso, $\ln(\psi_{tk't'k'}) = \gamma_{tt'} (\omega_k^{tt'} - \omega_{k+1}^{tt'}) (\omega_{k'}^{tt'} - \omega_{k'+1}^{tt'})$, caracterizando uma estrutura com mais parâmetros e que deve ser utilizada em casos nos quais a associação é mais complexa.

Ressalta-se que as estimativas das razões de chance não são afetadas pelos parâmetros β , mas estes variam conforme a estrutura escolhida (TOULOU MIS; AGRESTI; KATERI, 2013).

Para o exemplo descrito com 3 categorias e 3 instantes de tempo, assim como a Tabela 2.2, é possível construir a matriz de razões de chances da seguinte forma:

$$\begin{bmatrix} \mathbf{0} & \Psi_{12} & \Psi_{13} \\ \Psi_{21} & \mathbf{0} & \Psi_{23} \\ \Psi_{31} & \Psi_{32} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \begin{array}{cc|cc|cc} 0 & 0 & \psi_{1121} & \psi_{1122} & \psi_{1131} & \psi_{1132} \\ 0 & 0 & \psi_{1221} & \psi_{1222} & \psi_{1231} & \psi_{1232} \end{array} \\ \hline \begin{array}{cc|cc|cc} \psi_{2111} & \psi_{2112} & 0 & 0 & \psi_{2131} & \psi_{2132} \\ \psi_{2211} & \psi_{2212} & 0 & 0 & \psi_{2231} & \psi_{2232} \end{array} \\ \hline \begin{array}{cc|cc|cc} \psi_{3111} & \psi_{3112} & \psi_{3121} & \psi_{3122} & 0 & 0 \\ \psi_{3211} & \psi_{3212} & \psi_{3221} & \psi_{3222} & 0 & 0 \end{array} \end{bmatrix}.$$

Vale lembrar que a matriz é simétrica ($\Psi_{tt'} = \Psi_{t't}$), então os elementos ψ_{1121} e ψ_{2111} são iguais, por exemplo. A diferença entre as estruturas permutável e RC é que na primeira os blocos $\Psi_{tt'}$

são sempre iguais ($\Psi_{12} = \Psi_{13} = \Psi_{23}$); por outro lado, na estrutura RC esses blocos passam a diferir pois assume-se que as razões de chances mudam conforme o par de tempos analisado. No contexto desta metodologia a interpretação dessas razões de chances locais não é trivial e, muitas vezes, tem apenas o apelo de representar uma medida de associação entre as observações. Um exemplo de interpretação pode ser encontrado em Touloumis (2011).

Adicionalmente, Touloumis (2015) explica como realizar a seleção das variáveis que formarão o preditor linear do modelo. Sejam M_0 e M_1 dois modelos encaixados com vetores de parâmetros β_0 e $\beta_1 = (\beta'_0, \beta'_q)'$, respectivamente, e também uma matriz C de tal forma que $C\beta_1 = \beta_q$. Um teste para as hipóteses

$$\begin{cases} H_0 : & \beta_q = \mathbf{0} \\ H_a : & \beta_q \neq \mathbf{0} \end{cases}$$

é baseado na estatística de teste de Wald, sendo H_0 rejeitada ao nível de $(100 \times \alpha)\%$ de significância quando

$$(C\hat{\beta})' (nC\hat{\Sigma}C')^{-1} (C\hat{\beta}) \geq \chi^2_q(\alpha), \quad (2.31)$$

em que

$$\Sigma = \lim_{n \rightarrow \infty} n\Sigma_0^{-1}\Sigma_1\Sigma_0^{-1},$$

sendo as expressões de Σ_0 e Σ_1 detalhadas em Touloumis (2015). Ainda, $\hat{\beta}$ e $\hat{\Sigma}$ são estimados com base no modelo M_1 . Demais inferências acerca dos parâmetros β , como a obtenção de intervalos de confiança podem ser feitas sob o fato que assintoticamente, conhecendo-se $\hat{\alpha}$, $\sqrt{n}(\hat{\beta}_G - \beta) \sim N(\mathbf{0}, \Sigma)$.

2.3.2.2 Recursos computacionais

Esta metodologia foi implementada no pacote *multgee* (TOULOUMIS, 2015) do *software* R (R Core Team, 2014). O ajuste de um modelo com resposta nominal é feito por meio da função *nomLORgee*, enquanto que a função *ordLORgee* analisa respostas ordinais. Concentrando-se no caso nominal, a estrutura de associação para as razões de chances locais é determinada pelo argumento *LORstr* dentro da função *nomLORgee* e assume os valores *time.exch* ou *RC*, representando as duas opções descritas anteriormente na sua respectiva ordem.

No *software* SAS (SAS, 2011) não há procedimentos que façam o ajuste de modelos utilizando esta abordagem.

2.3.3 Modelos mistos

Enquanto as duas metodologias apresentadas anteriormente são extensões das EEG originalmente propostas por Liang e Zeger (1986), o modelo misto apresentado aqui não tem

grandes modificações sobre o que foi exposto na seção 2.2.2 - trata-se da inclusão de efeitos aleatórios ao modelo de logitos generalizados (seção 2.1.2.2).

2.3.3.1 Metodologia

De acordo com Hedeker (2003), o modelo de logitos generalizados com a inclusão de efeitos aleatórios é dado por

$$\ln \left(\frac{\pi_{itk}}{\pi_{itK}} \right) = \mathbf{x}'_{it} \boldsymbol{\beta}_k + \mathbf{z}'_{it} \mathbf{b}_{ki},$$

para $k = 1, \dots, K - 1$, $i = 1, \dots, n$ e $t = 1, \dots, T_i$. O vetor $\mathbf{x}_{it}$ tem dimensão $p \times 1$ sendo p o número de parâmetros de $\boldsymbol{\beta}_k$, $\mathbf{z}_{it}$ tem dimensão $q \times 1$ sendo q a quantidade de efeitos aleatórios de $\mathbf{b}_{ki}$. Novamente a suposição usual é que os efeitos aleatórios têm distribuição normal, ou seja, $\mathbf{b}_{ki} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_k)$. O processo de estimação dos parâmetros do modelo é análogo ao exposto na seção 2.2.2, sabendo-se que a expressão (2.22) no caso multinomial é denotada por

$$f_i(\mathbf{y}_i | \mathbf{b}_i) = \prod_{t=1}^{T_i} \prod_{k=1}^K \pi_{itk}^{y_{itk}}, \quad (2.32)$$

similarmente à expressão (2.29). Devido à presença dos efeitos aleatórios, as probabilidades π_{itk} em (2.32) são dadas por

$$\pi_{itk} = P[y_{it} = k | \mathbf{b}] = \frac{\exp(\mathbf{x}'_{it} \boldsymbol{\beta}_k + \mathbf{z}'_{it} \mathbf{b}_{ki})}{1 + \exp\left(\sum_{h=1}^{K-1} \mathbf{x}'_{it} \boldsymbol{\beta}_h + \mathbf{z}'_{it} \mathbf{b}_{hi}\right)}.$$

Outros resultados, como as razões de chances e intervalos de confiança para as mesmas e para as probabilidades de cada categoria descritos na seção 2.1.2.2 também passam a ser condicionados aos efeitos aleatórios, como será detalhado nos métodos (seção 3.2.2). A título de exemplo, um modelo de interceptos aleatórios considerando três categorias para a variável resposta é dado por

$$\ln \left(\frac{\pi_{it1}}{\pi_{it3}} \right) = \beta_{01} + b_{1i} + \beta_{11} t_{it}$$

e

$$\ln \left(\frac{\pi_{it2}}{\pi_{it3}} \right) = \beta_{02} + b_{2i} + \beta_{12} t_{it},$$

em que $b_{1i} \sim N(0, \sigma_1^2)$ e $b_{2i} \sim N(0, \sigma_2^2)$.

2.3.3.2 Recursos computacionais

No *software* R estão disponíveis os pacotes `mlogit` (CROISSANT, 2013) e `mixcat` (PAPAGEORGIOU; HINDE, 2012a), que possibilitam o ajuste de modelos mistos para dados politômicos nominais. O segundo pacote, no entanto, utiliza uma abordagem diferente da apresentada aqui, utilizando uma distribuição não-paramétrica para os efeitos aleatórios (maiores detalhes disponíveis em Papageorgiou e Hinde (2012b), Papageorgiou e Hinde (2012a) e Hartzel,

Agresti e Caffo (2001)). Quanto ao pacote `mlogit`, não fica claro na sua documentação os processos utilizados na resolução da expressão (2.23).

No *software* SAS (SAS, 2011) há o procedimento PROC GLIMMIX (SAS, 2008b), capaz de ajustar modelos com a referida estrutura. Várias opções quanto ao processo de estimação estão disponíveis para serem trabalhadas.

3 MATERIAL E MÉTODOS

3.1 Material

Os dados que ilustram este trabalho foram coletados em um experimento conduzido em esquema fatorial 2×2 casualizado em blocos, realizado pelo Departamento de Zootecnia da Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ), campus da Universidade de São Paulo localizado em Piracicaba/SP. Analisou-se o tipo de vegetação formada em uma área de pasto com capim elefante (*Pennisetum purpureum*), contando-se a frequência de touceiras, plantas invasoras e regiões com espaços vazios. As observações foram feitas utilizando-se uma linha de *nylon* nos piquetes utilizados no campo, em formato de “W”, totalizando sempre 40 pontos observados em cada piquete e observando-se o tipo de vegetação presente logo abaixo desses pontos. Os dois fatores a serem considerados são as condições de **pré** e **pós-pastejo**, ambos com 2 níveis. O **pré-pastejo** corresponde ao período anterior à entrada dos animais no campo (vacas leiteiras), sendo avaliado o percentual de luminosidade absorvida pelo capim elefante (95% ou 100%). O fator **pós-pastejo** corresponde à altura do capim remanescente quando os animais foram retirados do pasto (35 cm ou 45 cm). Por fim, as observações foram feitas de janeiro de 2011 a abril de 2012, período classificado da seguinte forma: verão 1 (janeiro a março/11), outono (abril a junho/11), inverno (julho a setembro/11), início da primavera (outubro a novembro/11), fim da primavera (novembro a dezembro/11) e verão 2 (janeiro a abril/12) (PEREIRA *et al.*, 2014). Na Figura 3.1 é apresentado um croqui do experimento.

Figura 3.1 – Croqui do experimento realizado

Como há sempre 40 pontos em cada piquete pode-se pensar nas frequências relativas de cada tipo de vegetação como proporções, sendo portanto conveniente analisar a variável

resposta deste experimento como politômica nominal com três níveis (touceiras, plantas invasoras e espaços vazios). A Tabela 3.1 apresenta a quantidade de observações presentes no estudo: levando-se em consideração que são 16 piquetes, cada um com 40 observações e há 6 estações do ano, há um total de $16 \times 40 \times 6 = 3.840$ observações. Porém, devido a uma geada, perderam-se as observações do piquete 13 no início da primavera e, portanto, têm-se $3.840 - 40 = 3.800$ pontos observados.

Tabela 3.1 – Quantidade de observações analisadas no experimento, por tratamento e estação do ano

Estação	Tratamento				Total
	95 : 35	95 : 45	100 : 35	100 : 45	
Verão 1	160	160	160	160	640
Outono	160	160	160	160	640
Inverno	160	160	160	160	640
Início da primavera	160	160	160	120	600
Fim da primavera	160	160	160	160	640
Verão 2	160	160	160	160	640
Total	960	960	960	920	3.800

3.2 Métodos

Nesta seção são apresentados os métodos utilizados na seleção dos modelos que serão aplicados ao conjunto de dados do experimento descrito. Primeiramente, são apresentadas as ferramentas para o ajuste via equações de estimação generalizadas e, posteriormente, aquelas utilizadas no ajuste via modelos mistos.

3.2.1 Ajuste via EEG

3.2.1.1 Seleção da estrutura de associação e das variáveis

O ajuste do modelo marginal é feito com base nos métodos descritos por Touloumis (2015) e Touloumis, Agresti e Kateri (2013):

Passo 1 - seleção da estrutura de associação: como as estimativas dos parâmetros de regressão dependem da estrutura de associação (permutável ou RC), primeiramente seleciona-se esta estrutura. Analisando-se as estimativas dos parâmetros intrínsecos (γ) sob a estrutura RC (mais complexa e que assume estes diferem entre si), se as estimativas não diferirem muito deve-se optar pela estrutura permutável; caso contrário, escolhe-se a estrutura RC. No pacote `multgee` isso é feito por meio da função `intrinsic.pars()`, que retorna tais estimativas.

Passo 2 - seleção das variáveis: como descrito ao final da seção 2.3.2.1, vários modelos encaixados podem ser propostos e comparados entre si. Neste trabalho, parte-se de um modelo

com todas as covariáveis (e interações) e depois são feitas várias simplificações do mesmo avaliando os efeitos da inclusão/exclusão das covariáveis. Ao todo são propostos 7 modelos, cujas estruturas adotadas no preditor linear são apresentadas na Tabela 3.2, em que a última coluna mostra o número de parâmetros presentes em cada logito. Observação: os fatores pré e pós-pastejo foram denominados de pré e pós, respectivamente, por conveniência. No pacote `multgee`, a função `waldts()` executa o teste de Wald descrito, cujos graus de liberdade são dados pelo dobro da diferença do número de parâmetros entre os modelos comparados, dado que são três categorias de resposta e, portanto, dois logitos.

Tabela 3.2 – Modelos encaixados propostos para o ajuste via EEG

Modelo	Preditor linear	Parâmetros
1	Bloco + Pré * Pós * Estação	27
2	Bloco + Pré + Pós + Estação + Pré × Pós + Pré × Estação + Pós × Estação	22
3	Bloco + Pré + Pós + Estação + Pré × Pós	12
4	Bloco + Pré + Pós + Estação + Pré × Pós + Pós × Estação	17
5	Bloco + Pré + Pós + Estação + Pré × Pós + Pré × Estação	17
6	Bloco + Pré + Pós + Estação + Pré × Estação	16
7	Bloco + Pré + Estação + Pré × Estação	15

Nota: no modelo 1, * simboliza a presença da interação tripla, das interações duplas e dos efeitos principais.

Sendo π_{itk} a probabilidade de o i -ésimo ponto amostral ($i = 1, \dots, 640$) ser classificado na k -ésima categoria ($k = 1, 2, 3$) no instante $t = (1, \dots, 6)$, o modelo é composto pelos logitos

$$\text{logito}[\pi_{it1}] = \ln\left(\frac{\pi_{it1}}{\pi_{it3}}\right) \quad \text{logito}[\pi_{it2}] = \ln\left(\frac{\pi_{it2}}{\pi_{it3}}\right).$$

Os índices $k = 1, 2, 3$ representam os tipos de vegetação touceira, planta invasora e espaço vazio, respectivamente, de modo que as interpretações são feitas em relação em relação à categoria “espaço vazio”. A seguir são detalhadas as expressões de cada um dos modelos apresentados na Tabela 3.2, utilizando variáveis *dummy* definidas na Tabela 3.3:

Modelo 1:

$$\begin{aligned} \ln\left(\frac{\pi_{itk}}{\pi_{it3}}\right) = & \underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{3i}}_{\text{pós}} + \\ & \underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{\text{estação}} + \underbrace{\beta_{11,k}X_{2i}X_{3i}}_{\text{pré} \times \text{pós}} + \\ & \underbrace{\beta_{12,k}X_{2i}X_{41t} + \dots + \beta_{16,k}X_{2i}X_{45t}}_{\text{pré} \times \text{estação}} + \underbrace{\beta_{17,k}X_{3i}X_{41t} + \dots + \beta_{21,k}X_{3i}X_{45t}}_{\text{pós} \times \text{estação}} + \\ & \underbrace{\beta_{22,k}X_{2i}X_{3i}X_{41t} + \dots + \beta_{26,k}X_{2i}X_{3i}X_{45t}}_{\text{pré} \times \text{pós} \times \text{estação}}. \end{aligned}$$

Modelo 2:

$$\ln \left(\frac{\pi_{itk}}{\pi_{it3}} \right) = \underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{3i}}_{\text{pós}} +$$

$$\underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{\text{estação}} + \underbrace{\beta_{11,k}X_{2i}X_{3i}}_{\text{pré} \times \text{pós}} +$$

$$\underbrace{\beta_{12,k}X_{2i}X_{41t} + \cdots + \beta_{16,k}X_{2i}X_{45t}}_{\text{pré} \times \text{estação}} + \underbrace{\beta_{17,k}X_{3i}X_{41t} + \cdots + \beta_{21,k}X_{3i}X_{45t}}_{\text{pós} \times \text{estação}}.$$

Modelo 3:

$$\ln \left(\frac{\pi_{itk}}{\pi_{it3}} \right) = \underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{3i}}_{\text{pós}} +$$

$$\underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{\text{estação}} + \underbrace{\beta_{11,k}X_{2i}X_{3i}}_{\text{pré} \times \text{pós}}.$$

Modelo 4:

$$\ln \left(\frac{\pi_{itk}}{\pi_{it3}} \right) = \underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{3i}}_{\text{pós}} +$$

$$\underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{\text{estação}} + \underbrace{\beta_{11,k}X_{2i}X_{3i}}_{\text{pré} \times \text{pós}} +$$

$$\underbrace{\beta_{12,k}X_{3i}X_{41t} + \cdots + \beta_{16,k}X_{3i}X_{45t}}_{\text{pós} \times \text{estação}}.$$

Modelo 5:

$$\ln \left(\frac{\pi_{itk}}{\pi_{it3}} \right) = \underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{3i}}_{\text{pós}} +$$

$$\underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{\text{estação}} + \underbrace{\beta_{11,k}X_{2i}X_{3i}}_{\text{pré} \times \text{pós}} +$$

$$\underbrace{\beta_{12,k}X_{2i}X_{41t} + \cdots + \beta_{16,k}X_{2i}X_{45t}}_{\text{pré} \times \text{estação}}.$$

Modelo 6:

$$\ln \left(\frac{\pi_{itk}}{\pi_{it3}} \right) = \underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{3i}}_{\text{pós}} +$$

$$\underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{\text{estação}} +$$

$$\underbrace{\beta_{11,k}X_{2i}X_{41t} + \cdots + \beta_{15,k}X_{2i}X_{45t}}_{\text{pré} \times \text{estação}}.$$

Modelo 7:

$$\ln \left(\frac{\pi_{itk}}{\pi_{it3}} \right) = \underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{41t} + \beta_{6k}X_{42t} + \beta_{7k}X_{43t} + \beta_{8k}X_{44t} + \beta_{9,k}X_{45t}}_{\text{estação}} + \underbrace{\beta_{10,k}X_{2i}X_{41t} + \cdots + \beta_{14,k}X_{2i}X_{45t}}_{\text{pré} \times \text{estação}}.$$

Tabela 3.3 – Codificações das variáveis *dummy* utilizadas no modelo ajustado via EEG

(a) Blocos				(b) Pré-pastejo		(c) Pós-pastejo	
Bloco	X_{11}	X_{12}	X_{13}	Pré-pastejo	X_2	Pós-pastejo	X_3
1	0	0	0				
2	1	0	0	95%	0	35 cm	0
3	0	1	0	100%	1	45 cm	1
4	0	0	1				

(d) Estações do ano					
Estação	X_{41}	X_{42}	X_{43}	X_{44}	X_{45}
Verão 1	0	0	0	0	0
Outono	1	0	0	0	0
Inverno	0	1	0	0	0
Início da primavera	0	0	1	0	0
Fim da primavera	0	0	0	1	0
Verão 2	0	0	0	0	1

3.2.1.2 Probabilidades preditas e intervalos de confiança

É de interesse neste tipo de análise a obtenção das probabilidades de ocorrência de cada categoria da variável resposta (neste caso, tipo de vegetação). Para tal, basicamente, aplica-se a expressão mostrada em (2.11). Considere, por exemplo, que o modelo 3 seja selecionado e que deseja-se saber a probabilidade de ocorrência de touceiras sob as condições: inverno, bloco 1, 100% de absorção de luminosidade no pré-pastejo e 35 cm de altura do pasto no pós-pastejo. Os valores estimados dos dois logitos do modelo são

$$\text{logito} [\hat{\pi}_{i31}] = \hat{\beta}_{01} + \hat{\beta}_{41} + \hat{\beta}_{71}$$

e

$$\text{logito} [\hat{\pi}_{i32}] = \hat{\beta}_{02} + \hat{\beta}_{42} + \hat{\beta}_{72},$$

respectivamente. Logo, pela expressão (2.11), a probabilidade estimada de ocorrência de touceira no i -ésimo ponto é dada por

$$\begin{aligned}\hat{\pi}_{i31} &= \frac{\exp [\text{logito} [\hat{\pi}_{i31}]]}{1 + \exp [\text{logito} [\hat{\pi}_{i31}]] + \exp [\text{logito} [\hat{\pi}_{i32}]]} \\ &= \frac{\exp (\hat{\beta}_{01} + \hat{\beta}_{41} + \hat{\beta}_{71})}{1 + \exp (\hat{\beta}_{01} + \hat{\beta}_{41} + \hat{\beta}_{71}) + \exp (\hat{\beta}_{02} + \hat{\beta}_{42} + \hat{\beta}_{72})}.\end{aligned}\quad (3.1)$$

Analogamente, a probabilidade de ocorrência de planta invasora no i -ésimo ponto é dada por

$$\begin{aligned}\hat{\pi}_{i32} &= \frac{\exp [\text{logito} [\hat{\pi}_{i32}]]}{1 + \exp [\text{logito} [\hat{\pi}_{i31}]] + \exp [\text{logito} [\hat{\pi}_{i32}]]} \\ &= \frac{\exp (\hat{\beta}_{02} + \hat{\beta}_{42} + \hat{\beta}_{72})}{1 + \exp (\hat{\beta}_{01} + \hat{\beta}_{41} + \hat{\beta}_{71}) + \exp (\hat{\beta}_{02} + \hat{\beta}_{42} + \hat{\beta}_{72})}\end{aligned}\quad (3.2)$$

e, como $\hat{\pi}_{it1} + \hat{\pi}_{it2} + \hat{\pi}_{it3} = 1$, a probabilidade de o i -ésimo ponto ser um espaço vazio é obtida fazendo-se $\hat{\pi}_{it3} = 1 - (\hat{\pi}_{it1} + \hat{\pi}_{it2})$. O mesmo raciocínio é empregado para a obtenção das probabilidades para as demais combinações das covariáveis presentes em qualquer um dos sete modelos apresentados.

É de interesse também a construção de intervalos de confiança para estas probabilidades, para se ter uma ideia da precisão das estimativas obtidas. O procedimento brevemente descrito na seção 2.1.2.1 é detalhado aqui pois os modelos apresentados têm diversas variáveis e/ou interações, merecendo atenção no cálculo dos intervalos.

Antes, contudo, é conveniente utilizar vetores e matrizes para facilitar o entendimento e também a implementação computacional utilizada (apresentada nos anexos deste trabalho), uma vez que o pacote `multgee` não disponibiliza estes resultados diretamente. Sejam β_1 e β_2 dois vetores contendo os parâmetros dos dois logitos de qualquer um dos modelos apresentados, respectivamente. Considere novamente, como ilustração, o modelo 3; estes vetores são dados por

$$\beta_1 = [\beta_{01} \ \beta_{11} \ \beta_{21} \ \beta_{31} \ \beta_{41} \ \beta_{51} \ \beta_{61} \ \beta_{71} \ \beta_{81} \ \beta_{91} \ \beta_{10,1} \ \beta_{11,1}]'$$

e

$$\beta_2 = [\beta_{02} \ \beta_{12} \ \beta_{22} \ \beta_{32} \ \beta_{42} \ \beta_{52} \ \beta_{62} \ \beta_{72} \ \beta_{82} \ \beta_{92} \ \beta_{10,2} \ \beta_{11,2}]'.$$

Defina-se ainda um vetor C contendo valores 0 e 1 atribuídos às posições relativas aos parâmetros de interesse para uma certa combinação das covariáveis do modelo em estudo. Para as condições citadas no início desta seção (bloco 1, pré-pastejo igual a 100%, pós-pastejo igual a 35 cm e inverno como estação do ano), os respectivos parâmetros do modelo 3 são β_{0k} , β_{4k} e β_{7k} . O vetor C é definido então como

$$C = [1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0]'$$

O primeiro passo da construção dos intervalos de confiança para as probabilidades é a obtenção das estimativas de ambos os logitos do modelo, que utilizando o vetor $\mathbf{C}$ são dadas por

$$\text{logito} [\hat{\pi}_{i3k}] = \mathbf{C}' \boldsymbol{\beta}_k = [1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0] \begin{bmatrix} \hat{\beta}_{0k} \\ \hat{\beta}_{1k} \\ \hat{\beta}_{2k} \\ \hat{\beta}_{3k} \\ \hat{\beta}_{4k} \\ \hat{\beta}_{5k} \\ \hat{\beta}_{6k} \\ \hat{\beta}_{7k} \\ \hat{\beta}_{8k} \\ \hat{\beta}_{9k} \\ \hat{\beta}_{10,k} \\ \hat{\beta}_{11,k} \end{bmatrix} = \hat{\beta}_{0k} + \hat{\beta}_{4k} + \hat{\beta}_{7k}.$$

O segundo passo é a obtenção da variância associada a cada logito. No exemplo, tem-se

$$\begin{aligned} \hat{\text{Var}} [\text{logito}[\hat{\pi}_{i3k}]] &= \hat{\text{Var}} [\hat{\beta}_{0k}] + \hat{\text{Var}} [\hat{\beta}_{4k}] + \hat{\text{Var}} [\hat{\beta}_{7k}] + \\ &\quad 2\hat{\text{cov}} [\hat{\beta}_{0k}, \hat{\beta}_{4k}] + 2\hat{\text{cov}} [\hat{\beta}_{0k}, \hat{\beta}_{7k}] + 2\hat{\text{cov}} [\hat{\beta}_{4k}, \hat{\beta}_{7k}] \end{aligned}$$

que, matricialmente, é simbolizado por

$$\hat{\text{Var}} [\text{logito}[\hat{\pi}_{i3k}]] = \mathbf{C}' \hat{\text{Var}} [\hat{\boldsymbol{\beta}}_k] \mathbf{C},$$

em que $\hat{\text{Var}} [\hat{\boldsymbol{\beta}}_k]$ é a matriz de variâncias e covariâncias das estimativas dos parâmetros $\boldsymbol{\beta}_k$, usualmente disponível nos *softwares*. Essa matriz tem a dimensão $p \times p$, em que p é o número que parâmetros do modelo em questão, e no exemplo é dada por

$$\begin{bmatrix} \hat{\text{Var}} [\hat{\beta}_{0k}] & \hat{\text{cov}} [\hat{\beta}_{0k}, \hat{\beta}_{1k}] & \hat{\text{cov}} [\hat{\beta}_{0k}, \hat{\beta}_{2k}] & \dots & \hat{\text{cov}} [\hat{\beta}_{0k}, \hat{\beta}_{11,k}] \\ \hat{\text{cov}} [\hat{\beta}_{1k}, \hat{\beta}_{0k}] & \hat{\text{Var}} [\hat{\beta}_{1k}] & \hat{\text{cov}} [\hat{\beta}_{1k}, \hat{\beta}_{2k}] & \dots & \hat{\text{cov}} [\hat{\beta}_{1k}, \hat{\beta}_{11,k}] \\ \hat{\text{cov}} [\hat{\beta}_{2k}, \hat{\beta}_{0k}] & \hat{\text{cov}} [\hat{\beta}_{2k}, \hat{\beta}_{1k}] & \hat{\text{Var}} [\hat{\beta}_{2k}] & \dots & \hat{\text{cov}} [\hat{\beta}_{2k}, \hat{\beta}_{11,k}] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\text{cov}} [\hat{\beta}_{11,k}, \hat{\beta}_{0k}] & \hat{\text{cov}} [\hat{\beta}_{11,k}, \hat{\beta}_{1k}] & \hat{\text{cov}} [\hat{\beta}_{11,k}, \hat{\beta}_{2k}] & \dots & \hat{\text{Var}} [\hat{\beta}_{11,k}] \end{bmatrix}_{12 \times 12}.$$

Com isso, pode-se construir um intervalo de confiança para os logitos:

$$\text{logito} [\hat{\pi}_{i3k}] \pm z_{(1-\frac{\alpha}{2})} \times \sqrt{\hat{\text{Var}} [\text{logito} [\hat{\pi}_{i3k}]]}. \quad (3.3)$$

Finalizando, os intervalos de confiança para as probabilidades de ocorrência são obtidos aplicando-se as expressões (3.1) e (3.2), porém agora com os limites inferior e superior dos intervalos de confiança para os logitos dados por (3.3).

3.2.1.3 Avaliação da qualidade do ajuste

Uma vez ajustado o modelo, é preciso avaliar a qualidade do ajuste aos dados (se há valores discrepantes, se o ajuste é satisfatório, por exemplo). A fim de avaliar como o modelo ajusta-se aos dados são comparados os valores preditos e observados, plotando-se as curvas e obtendo a correlação entre eles; quanto maior a correlação entre eles melhor é o ajuste do modelo.

Devem ser analisados também os resíduos do modelo. Aqui são avaliados os resíduos brutos, observando-se se há padrões, tendências ou valores discrepantes quando plotados *versus* a ordem das observações.

3.2.1.4 Interpretações dos resultados

Por se tratar de uma extensão dos modelos de logitos generalizados, a interpretação dos parâmetros é feita em função de razões de chances entre os níveis das variáveis comparando principalmente os níveis dos fatores pré e pós-pastejo, obtidas segundo o que foi exposto na seção 2.1.2.2. Nos modelos propostos, atenção especial deve ser dada às razões de chances envolvendo variáveis que tenham interações com outras. Por fim, mas de objetivo secundário neste tipo de modelo, são interpretadas também as razões de chances locais que descrevem a dependência entre as observações.

3.2.2 Ajuste via modelos mistos

3.2.2.1 Seleção do modelo

O modelo misto a ser ajustado parte do preditor linear escolhido no modelo ajustado via EEG (ver Tabela 3.2), incluindo-se efeitos aleatórios para considerar a dependência entre observações de um mesmo indivíduo. O único efeito aleatório considerado aqui é para os interceptos, por considerar-se que não há razão para adicionar efeitos aleatórios nas demais covariáveis do modelo (o que complicaria também as interpretações, uma vez que devem-se fixar os efeitos aleatórios como descrito na seção 2.2.3). Em termos práticos, ajustando-se um modelo de interceptos aleatórios considera-se que os 640 pontos (“indivíduos”) têm uma grande variabilidade entre si e que as demais covariáveis exercem a mesma influência sobre todos os pontos (não há motivos para acreditar, por exemplo, que as estações do ano influenciem de forma distinta os pontos). Por exemplo, se o modelo 2 for selecionado, o modelo que antes era expresso por

$$\ln \left(\frac{\pi_{itk}}{\pi_{it3}} \right) = \underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{3i}}_{\text{pós}} +$$

$$\underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{\text{estação}} + \underbrace{\beta_{11,k}X_{2i}X_{3i}}_{\text{pré} \times \text{pós}} +$$

$$\underbrace{\beta_{12,k}X_{2i}X_{41t} + \cdots + \beta_{16,k}X_{2i}X_{45t}}_{\text{pré} \times \text{estação}} + \underbrace{\beta_{17,k}X_{3i}X_{41t} + \cdots + \beta_{21,k}X_{3i}X_{45t}}_{\text{pós} \times \text{estação}}.$$

com a inclusão dos efeitos aleatórios passa a ser

$$\ln\left(\frac{\pi_{itk}}{\pi_{it3}}\right) = \left(\underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{b_{ki}}_{\text{efeito aleatório}} \right) + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{3i}}_{\text{pós}} +$$

$$\underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{\text{estação}} + \underbrace{\beta_{11,k}X_{2i}X_{3i}}_{\text{pré} \times \text{pós}} +$$

$$\underbrace{\beta_{12,k}X_{2i}X_{41t} + \dots + \beta_{16,k}X_{2i}X_{45t}}_{\text{pré} \times \text{estação}} + \underbrace{\beta_{17,k}X_{3i}X_{41t} + \dots + \beta_{21,k}X_{3i}X_{45t}}_{\text{pós} \times \text{estação}}.$$

Qualquer que seja o modelo selecionado dentre os sete propostos na Tabela 3.2, os efeitos aleatórios b_{ki} são diferentes conforme o logito em questão. Ainda, assume-se neste trabalho que estes efeitos aleatórios têm distribuição normal de modo que

$$\begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right),$$

em que σ_{12} é a covariância entre os efeitos aleatórios.

A fim de avaliar quão forte é a dependência entre as observações de cada ponto é ajustado ainda um modelo sem efeitos aleatórios, isto é, um modelo apenas de efeitos fixos e sem nenhuma estrutura de associação/dependência. A comparação entre este e o modelo de interceptos aleatórios é feita utilizando-se o critério AIC (AKAIKE, 1974) e o teste da razão de verossimilhanças (TRV). Este teste avalia, sob H_0 , se $\sigma_1^2 = 0$ e $\sigma_2^2 = 0$; usualmente, a distribuição assintótica relacionada ao teste é uma qui-quadrado com 1 grau de liberdade. Porém, como o parâmetro σ^2 está localizado na fronteira do espaço paramétrico, é prudente considerar uma mistura de qui-quadrados como distribuição $\left(\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2\right)$ (SILVA, 2013).

O ajuste do modelo misto é feito pelo PROC GLIMMIX (SAS, 2008b) do SAS (SAS, 2011), utilizando a quadratura Gaussiana adaptativa no processo de estimação. O modelo sem efeitos aleatórios é ajustado também pelo PROC GLIMMIX.

3.2.2.2 Probabilidades preditas e intervalos de confiança

As probabilidades de ocorrência, bem como os respectivos intervalos de confiança para elas, podem ser obtidas utilizando a mesma metodologia para o ajuste via EEG detalhada na seção 3.2.1.2. A diferença é que, no modelo misto, cada ponto amostral tem uma probabilidade distinta devido aos efeitos aleatórios. Por exemplo, sob o modelo 2 com a inclusão dos efeitos aleatórios, os valores dos logits para o quinto ponto nas mesmas condições do exemplo citado na seção 3.2.1.2 são

$$\text{logito}[\hat{\pi}_{531}] = (\hat{\beta}_{01} + \hat{b}_{15}) + \hat{\beta}_{41} + \hat{\beta}_{71} + \hat{\beta}_{13,1}$$

e

$$\text{logito}[\hat{\pi}_{532}] = (\hat{\beta}_{02} + \hat{b}_{25}) + \hat{\beta}_{42} + \hat{\beta}_{72} + \hat{\beta}_{13,2},$$

em que $\hat{b}_{ki}$ são as predições dos efeitos aleatórios. As probabilidades de ocorrência de touceira e planta invasora são determinadas, respectivamente, por

$$\begin{aligned}\hat{\pi}_{531} &= \frac{\exp [\text{logito} [\hat{\pi}_{531}]]}{1 + \exp [\text{logito} [\hat{\pi}_{531}]] + \exp [\text{logito} [\hat{\pi}_{532}]]} \\ &= \frac{\exp [(\hat{\beta}_{01} + \hat{b}_{15}) + \hat{\beta}_{41} + \hat{\beta}_{71} + \hat{\beta}_{13,1}]}{1 + \exp [(\hat{\beta}_{01} + \hat{b}_{15}) + \hat{\beta}_{41} + \hat{\beta}_{71} + \hat{\beta}_{13,1}] + \exp [(\hat{\beta}_{02} + \hat{b}_{25}) + \hat{\beta}_{42} + \hat{\beta}_{72} + \hat{\beta}_{13,2}]}\end{aligned}$$

e

$$\begin{aligned}\hat{\pi}_{532} &= \frac{\exp [\text{logito} [\hat{\pi}_{532}]]}{1 + \exp [\text{logito} [\hat{\pi}_{531}]] + \exp [\text{logito} [\hat{\pi}_{532}]]} \\ &= \frac{\exp [(\hat{\beta}_{02} + \hat{b}_{25}) + \hat{\beta}_{42} + \hat{\beta}_{72} + \hat{\beta}_{13,2}]}{1 + \exp [(\hat{\beta}_{01} + \hat{b}_{15}) + \hat{\beta}_{41} + \hat{\beta}_{71} + \hat{\beta}_{13,1}] + \exp [(\hat{\beta}_{02} + \hat{b}_{25}) + \hat{\beta}_{42} + \hat{\beta}_{72} + \hat{\beta}_{13,2}]}\end{aligned}$$

Analogamente, cada “indivíduo” (ponto amostral) tem uma variância associada a seu logito. Seguindo com o exemplo, a variância do k -ésimo logito é dada por

$$\begin{aligned}\text{Var} [\text{logito} [\hat{\pi}_{53k}]] &= \text{Var} [\hat{\beta}_{0k}] + \text{Var} [\hat{b}_{k5}] + \text{Var} [\hat{\beta}_{4k}] + \text{Var} [\hat{\beta}_{7k}] + \text{Var} [\hat{\beta}_{13,k}] + \\ &\quad 2\text{cov} [\hat{\beta}_{0k}, \hat{\beta}_{4k}] + 2\text{cov} [\hat{\beta}_{0k}, \hat{\beta}_{7k}] + 2\text{cov} [\hat{\beta}_{0k}, \hat{\beta}_{13,k}] + \\ &\quad 2\text{cov} [\hat{\beta}_{4k}, \hat{\beta}_{7k}] + 2\text{cov} [\hat{\beta}_{4k}, \hat{\beta}_{13,k}] + 2\text{cov} [\hat{\beta}_{7k}, \hat{\beta}_{13,k}],\end{aligned}$$

pois $\text{cov} [\hat{b}_{ki}, \hat{\beta}_{rk}] = 0$, em que r denota o índice dos parâmetros de regressão. Logo, cada indivíduo tem um intervalo de confiança para a probabilidade de ocorrência, obtido conforme demonstrado na seção 3.2.1.2. Um intervalo de confiança geral para todos os indivíduos pode ser obtido fazendo-se a média dos intervalos de cada indivíduo i ($i = 1, \dots, 640$).

3.2.2.3 Avaliação da qualidade do ajuste

A qualidade do ajuste será avaliada de forma análoga ao modelo ajustado via EEG, por meio de comparações entre os valores observados e preditos (curvas médias). Como descrito na seção 2.2.3, não há uma expressão analítica para o efeito médio em um MLGM; considerando-se o fato que o tamanho amostral é grande, as curvas médias ajustadas serão feitas utilizando-se a média das curvas individuais de cada ponto amostral.

3.2.2.4 Interpretações dos resultados

A interpretação dos parâmetros, naturalmente, também pode ser feita por meio de razões de chances comparando os níveis das variáveis explicativas assim como no modelo ajustado via EEG. Contudo, foi constatado na seção 2.2.3 que tal interpretação só faz sentido para covariáveis que variem ao longo do tempo para um mesmo indivíduo (ponto amostral, nesse caso). Os fatores pré e pós-pastejo são invariantes no tempo, ou seja, cada ponto recebe sempre o mesmo tratamento, o que impossibilita a interpretação destes parâmetros no modelo ajustado como um MLGM. Assim, as únicas interpretações que podem ser feitas são acerca dos parâmetros referentes às estações do ano, caso essa variável seja significativa.

4 RESULTADOS E DISCUSSÃO

4.1 Análise exploratória

Na Tabela 4.1 são apresentadas as proporções médias observadas de cada tipo de vegetação, e seus respectivos desvios padrões. Nota-se a predominância de touceiras na maioria dos tratamentos e estações do ano, mas também que quando há 100% de luminosidade absorvida (nível 2 do tratamento pré-pastejo), há uma queda no percentual de touceiras e aumento de espaços vazios no final da primavera e no segundo verão. Essas proporções são mostradas na Figura 4.1.

Tabela 4.1 – Proporções médias observadas (p) de touceira, planta invasora e espaço vazio em cada tratamento e estação do ano, e seus respectivos desvios padrões (d.p.)

Estação	Tratamento	Touceira		Planta invasora		Espaço vazio	
		p	d.p.	p	d.p.	p	d.p.
Verão 1	100 : 35	48,8%	6,0%	8,1%	8,0%	43,1%	12,6%
Verão 1	100 : 45	64,4%	10,5%	8,1%	1,3%	27,5%	10,6%
Verão 1	95 : 35	48,8%	8,8%	12,5%	11,4%	38,8%	4,3%
Verão 1	95 : 45	56,9%	8,0%	9,4%	6,9%	33,8%	2,5%
Outono	100 : 35	55,0%	4,1%	8,8%	6,6%	36,3%	7,2%
Outono	100 : 45	56,3%	6,0%	3,1%	1,3%	40,6%	5,5%
Outono	95 : 35	56,9%	13,6%	14,4%	8,3%	28,8%	6,6%
Outono	95 : 45	58,8%	8,3%	8,8%	4,3%	32,5%	9,6%
Inverno	100 : 35	55,0%	9,4%	5,0%	4,6%	40,0%	7,9%
Inverno	100 : 45	56,3%	4,3%	3,8%	4,3%	40,0%	6,5%
Inverno	95 : 35	59,4%	3,2%	11,3%	3,2%	29,4%	3,2%
Inverno	95 : 45	58,8%	4,8%	3,8%	1,4%	37,5%	5,4%
Início da primavera	100 : 35	51,9%	4,3%	5,6%	1,3%	42,5%	4,6%
Início da primavera	100 : 45	55,0%	2,5%	3,3%	3,8%	41,7%	6,3%
Início da primavera	95 : 35	65,0%	5,8%	4,4%	3,8%	30,6%	2,4%
Início da primavera	95 : 45	65,0%	3,5%	3,8%	4,3%	31,3%	3,2%
Fim da primavera	100 : 35	48,1%	6,9%	6,3%	8,3%	45,6%	7,5%
Fim da primavera	100 : 45	46,9%	6,6%	9,4%	5,5%	43,8%	10,5%
Fim da primavera	95 : 35	59,4%	4,7%	8,8%	4,3%	31,9%	7,2%
Fim da primavera	95 : 45	58,8%	3,2%	4,4%	4,3%	36,9%	7,2%
Verão 2	100 : 35	45,0%	5,8%	4,4%	5,5%	50,6%	4,7%
Verão 2	100 : 45	48,1%	7,2%	3,8%	4,3%	48,1%	3,8%
Verão 2	95 : 35	59,4%	10,1%	10,0%	10,6%	30,6%	1,3%
Verão 2	95 : 45	58,1%	7,5%	8,8%	9,2%	33,1%	12,3%
Total		55,7%	8,3%	7,1%	6,1%	37,2%	8,8%

Figura 4.1 – Proporções médias observadas de cada tipo de vegetação ao longo das estações do ano, conforme os tratamentos utilizados

4.2 Modelo ajustado via EEG

4.2.1 Seleção e formulação do modelo

Segundo o exposto na seção 3.2.1, primeiramente seleciona-se a estrutura de associação para as razões de chances locais. Analisando-se as estimativas dos parâmetros intrínsecos γ segundo a estrutura RC (Tabela 4.2), nota-se que os valores variam de $-0,0833$ a $0,8438$. Dada esta grande variação e por haver valores positivos e negativos, optou-se por adotar a estrutura RC no modelo a ser ajustado.

Tabela 4.2 – Estimativas dos parâmetros intrínsecos do ajuste via EEG, utilizando a estrutura RC

Parâmetro	Estimativa
γ_{12}	0,1689
γ_{13}	0,2853
γ_{14}	-0,0833
γ_{15}	0,0050
γ_{16}	0,5885
γ_{23}	0,6207
γ_{24}	0,4221
γ_{25}	0,8438
γ_{26}	0,7735
γ_{34}	0,8018
γ_{35}	0,3141
γ_{36}	0,0274
γ_{45}	0,4072
γ_{46}	0,8311
γ_{56}	0,0891

Para a seleção das covariáveis do preditor linear foram comparados os modelos expostos na Tabela 3.2. Na Tabela 4.3 são apresentados os p -valores resultantes do teste de hipóteses definido em (2.31), na qual os graus de liberdade dos testes aparecem multiplicados por 2 pois o modelo é composto por 2 logitos. Na primeira coluna é apresentada a comparação dos modelos dois a dois, sempre mostrando primeiramente o modelo com mais parâmetros, representado pela hipótese alternativa do teste de Wald. Ao nível de 5% de significância conclui-se que:

- i) Na primeira comparação, entre os modelos 1 e 2, testa-se a necessidade de incluir a interação tripla entre os fatores e a estação do ano no modelo; pelo p -valor de 0,7089 conclui-se que esta não é necessária e escolhe-se o modelo 2 (sem interação tripla);
- ii) Comparando os modelos 2 e 3 testam-se se as duas interações entre os fatores e a estação do ano poderiam ser retiradas ao mesmo tempo do modelo. Pela significância observada nota-se que pelo menos uma das interações deve permanecer;
- iii) Nas duas comparações seguintes, do modelo 2 *versus* os modelos 4 e 5, avaliam-se as retiradas das interações entre os dois fatores e a estação do ano, uma de cada vez. Conclui-se que apenas a interação que inclui o pré-pastejo deve permanecer no modelo;
- iv) Após isso, comparando os modelos 5 e 6, nota-se que a interação entre os fatores pré e pós-pastejo não é significativa;
- v) Por fim, testa-se a significância do fator pós-pastejo (modelos 6 e 7), que permanece no modelo. Portanto, o modelo 6 é selecionado ao final do processo.

Tabela 4.3 – Resultado do teste de Wald nas comparações dos modelos encaixados ajustados via EEG

Comparação	Graus de liberdade	p-valor	Escolha
Modelo 1 × modelo 2	10	0,7089	modelo 2
Modelo 2 × modelo 3	20	0,0173	modelo 2
Modelo 2 × modelo 4	10	0,0174	modelo 2
Modelo 2 × modelo 5	10	0,1608	modelo 5
Modelo 5 × modelo 6	2	0,1700	modelo 6
Modelo 6 × modelo 7	2	0,0203	modelo 6

A expressão do modelo 6, conforme descrito na seção 3.2.1, é dada por

$$\begin{aligned}
 \ln \left(\frac{\pi_{itk}}{\pi_{it3}} \right) = & \underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{3i}}_{\text{pós}} + \\
 & \underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{\text{estação}} + \\
 & \underbrace{\beta_{11,k}X_{2i}X_{41t} + \cdots + \beta_{15,k}X_{2i}X_{45t}}_{\text{pré} \times \text{estação}},
 \end{aligned} \tag{4.1}$$

para $k = 1, 2$. Se houver interesse em comparar as categorias touceira e planta invasora ($k = 1, 2$) basta fazer a diferença entre os dois respectivos logitos, pois

$$\ln \left(\frac{\pi_{it1}}{\pi_{it3}} \right) - \ln \left(\frac{\pi_{it2}}{\pi_{it3}} \right) = \ln \left(\frac{\pi_{it1}}{\pi_{it2}} \right).$$

As estimativas dos parâmetros do modelo selecionado são apresentadas na Tabela 4.4 e serão interpretadas na sequência do texto. Os erros padrões são indicados entre parênteses ao lado de cada estimativa, e os efeitos significativos a 5% são sinalizados com um asterisco.

Embora as estimativas da comparação entre as categorias ($k = 1, 2$), (touceira × planta invasora), possam ser obtidas pela diferença entre os logitos ajustados pelo modelo (4.1), para se chegar aos respectivos erros padrões é necessário ajustar um modelo trocando-se a ordem das categorias (mantendo-se o mesmo preditor linear). Como descrito na seção 3.2.1.2, estes erros padrões são necessários para a obtenção dos intervalos de confiança para as probabilidades das categorias. Ajustando um modelo com as categorias na ordem (espaço vazio, planta invasora, touceira), o logito que compara as categorias citadas tem suas estimativas dadas pela Tabela 4.5. Neste “segundo” modelo, as estimativas dos parâmetros do logito que compara (espaço vazio × touceira) são, logicamente, os valores da Tabela 4.4a multiplicados por (-1) , com os mesmos erros padrões.

Tabela 4.4 – Estimativas dos parâmetros e erros padrões (entre parênteses) do modelo 6, ajustado via EEG

(a) Primeiro logito - touceira $\times$ espaço vazio		
Parâmetro		Estimativa
Intercepto	β_{01}	0,265 (0,137)
Bloco 2	β_{11}	0,174 (0,092)
Bloco 3	β_{21}	0,050 (0,089)
Bloco 4	β_{31}	0,118 (0,096)
Pré (100%)	β_{41}	0,095 (0,170)
Pós (45 cm)	β_{51}	0,052 (0,066)
Outono	β_{61}	0,259 (0,174)
Inverno	β_{71}	0,193 (0,170)
Início da primavera	β_{81}^*	0,366 (0,176)
Fim da primavera	β_{91}	0,165 (0,181)
Verão 2	$\beta_{10,1}$	0,235 (0,167)
Pré (100%) : Outono	$\beta_{11,1}$	-0,361 (0,246)
Pré (100%) : Inverno	$\beta_{12,1}$	-0,335 (0,240)
Pré (100%) : Início da primavera	$\beta_{13,1}^*$	-0,597 (0,252)
Pré (100%) : Fim da primavera	$\beta_{14,1}^*$	-0,576 (0,246)
Pré (100%) : Verão 2	$\beta_{15,1}^*$	-0,766 (0,238)

(b) Segundo logito - planta invasora $\times$ espaço vazio		
Parâmetro		Estimativa
Intercepto	β_{02}^*	-0,764 (0,220)
Bloco 2	β_{12}	0,137 (0,176)
Bloco 3	β_{22}^*	-1,195 (0,233)
Bloco 4	β_{32}^*	-0,442 (0,191)
Pré (100%)	β_{42}	-0,282 (0,294)
Pós (45 cm)	β_{52}^*	-0,329 (0,142)
Outono	β_{62}	0,220 (0,275)
Inverno	β_{72}	-0,314 (0,301)
Início da primavera	β_{82}^*	-0,851 (0,349)
Fim da primavera	β_{92}	-0,469 (0,301)
Verão 2	$\beta_{10,2}$	-0,042 (0,272)
Pré (100%) : Outono	$\beta_{11,2}$	-0,601 (0,428)
Pré (100%) : Inverno	$\beta_{12,2}$	-0,422 (0,460)
Pré (100%) : Início da primavera	$\beta_{13,2}$	0,075 (0,516)
Pré (100%) : Fim da primavera	$\beta_{14,2}$	0,194 (0,444)
Pré (100%) : Verão 2	$\beta_{15,2}^*$	-0,990 (0,468)

Tabela 4.5 – Estimativas dos parâmetros e erros padrões (entre parênteses) do modelo ajustado 6 via EEG, para a comparação touceira × planta invasora

Parâmetro		Estimativa
Intercepto	β_{01}^*	1,029 (0,209)
Bloco 2	β_{11}	0,037 (0,166)
Bloco 3	β_{21}^*	1,245 (0,230)
Bloco 4	β_{31}^*	0,560 (0,185)
Pré (100%)	β_{41}	0,377 (0,282)
Pós (45 cm)	β_{51}^*	0,380 (0,137)
Outono	β_{61}	0,039 (0,268)
Inverno	β_{71}	0,507 (0,290)
Início da primavera	β_{81}^*	1,217 (0,342)
Fim da primavera	β_{91}^*	0,634 (0,294)
Verão 2	$\beta_{10,1}$	0,278 (0,254)
Pré (100%) : Outono	$\beta_{11,1}$	0,240 (0,416)
Pré (100%) : Inverno	$\beta_{12,1}$	0,087 (0,450)
Pré (100%) : Início da primavera	$\beta_{13,1}$	-0,672 (0,505)
Pré (100%) : Fim da primavera	$\beta_{14,1}^*$	-0,769 (0,434)
Pré (100%) : Verão 2	$\beta_{15,1}$	0,225 (0,448)

4.2.2 Qualidade do ajuste

Para assegurar a qualidade do ajuste, primeiramente são apresentadas as curvas dos valores observados e estimados pelo modelo 6 na Figura 4.2 de acordo com os dois fatores e as estações do ano, e verifica-se que o ajuste aos dados é bastante satisfatório (até mesmo pela grande quantidade de parâmetros do modelo). Seria possível detalhar este gráfico também por blocos mas optou-se por mostrar os valores médios dos 4 blocos a fim de reduzir os resultados a um nível mais prático.

Dando sequência à avaliação da qualidade do ajuste, a Figura 4.3 mostra um gráfico de dispersão entre os valores observados e preditos, bem como os coeficientes de correlação linear de Pearson associados. Nota-se que, quando se analisam conjuntamente todos os pontos dos 3 tipos de vegetação, a correlação é bastante alta (0,9643); entretanto, analisando-se cada um dos 3 tipos isoladamente a correlação decai para um nível moderado. Isto pode ser explicado pelo fato de que em alguns blocos as curvas observadas e estimadas distanciam-se em algumas estações do ano. Porém, como já mostra a Figura 4.2, valores médios dos 4 blocos ficam bastante próximos às curvas e o ajuste é considerado satisfatório.

Na Figura 4.4 são mostrados os resíduos brutos do modelo ajustado, na qual não se observa nenhum padrão ou indício de má qualidade no ajuste, nem pontos discrepantes. Os resíduos estão compreendidos no intervalo de -1 a 1, e as linhas vermelhas representam as médias dos resíduos de cada logito, praticamente nulas.

Figura 4.2 – Comparação das proporções observadas e previstas pelo modelo 6, ajustado via EEG

Figura 4.3 – Correlação entre os valores observados e previstos pelo modelo 6, ajustado via EEG

Figura 4.4 – Resíduos do modelo 6 ajustado via EEG, de acordo com cada um dos logits

4.2.3 Interpretações e discussão

Antes de discutirem-se os resultados e suas interpretações deve-se lembrar que este modelo, ajustado via EEG, é um modelo marginal e, portanto, as inferências são feitas em relação à média das observações que compartilham as mesmas características.

Como pode ser notado na Figura 4.2, a probabilidade de ocorrência de plantas invasoras é aproximadamente a mesma quando se comparam os tratamentos, sendo sempre a menor das três. Em se tratando do **pré-pastejo**, a probabilidade de ocorrência de touceiras é maior para 95% de absorção de luminosidade. A interação com a estação do ano faz com que, para 95% de absorção, essa probabilidade aumente ao longo do tempo e depois caia um pouco; em contrapartida, para 100% de absorção, a probabilidade de touceiras sempre cai ao longo do tempo. A presença de interação pode ser observada também para a probabilidade de haver espaços vazios, que só aumenta ao longo do tempo para 100% de absorção enquanto tem um comportamento alternado para 95%.

Quanto ao **pós-pastejo**, nota-se claramente por que não houve interação com a estação do ano. Para ambos os níveis (35 e 45 cm), os comportamentos das probabilidades dos 3 tipos de vegetação seguem o mesmo padrão. Embora essa variável tenha sido significativa, as diferenças entre seus níveis não são tão evidentes quanto no pré-pastejo.

As probabilidades de ocorrência de cada tipo de vegetação podem ser obtidas utilizando a expressão (2.11). Por se tratar de um modelo marginal os pontos pertencentes a um mesmo piquete têm a mesma probabilidade de ocorrência de cada tipo de vegetação. Como exemplo, a

probabilidade de haver uma planta invasora em qualquer ponto i de um piquete do bloco 3 com 100% de absorção de luminosidade no pré-pastejo e 45cm de altura no pós-pastejo, no verão 1, é

$$\hat{\pi}_{i12} = \frac{\exp(\hat{\beta}_{02} + \hat{\beta}_{22} + \hat{\beta}_{42} + \hat{\beta}_{52})}{1 + \exp(\hat{\beta}_{01} + \hat{\beta}_{21} + \hat{\beta}_{41} + \hat{\beta}_{51}) + \exp(\hat{\beta}_{02} + \hat{\beta}_{22} + \hat{\beta}_{42} + \hat{\beta}_{52})} = 0,0287.$$

A probabilidade estimada de ocorrer touceira ($\hat{\pi}_{i11}$) é obtida analogamente, e a probabilidade de ser um espaço vazio é obtida fazendo-se $\hat{\pi}_{i13} = 1 - (\hat{\pi}_{i11} + \hat{\pi}_{i12})$. Intervalos de confiança para estas probabilidades podem ser obtidos conforme o exposto na seção 3.2.1.2. Para a probabilidade descrita acima, o intervalo é obtido pelos seguintes passos:

i) O valor do logito é

$$\text{logito}[\hat{\pi}_{i12}] = \hat{\beta}_{02} + \hat{\beta}_{22} + \hat{\beta}_{42} + \hat{\beta}_{52} = -2,570.$$

ii) A variância deste logito é

$$\begin{aligned} \hat{\text{Var}}[\text{logito}[\hat{\pi}_{i12}]] &= \hat{\text{Var}}[\hat{\beta}_{02}] + \hat{\text{Var}}[\hat{\beta}_{22}] + \hat{\text{Var}}[\hat{\beta}_{42}] + \hat{\text{Var}}[\hat{\beta}_{52}] + \\ &2\hat{\text{cov}}[\hat{\beta}_{02}, \hat{\beta}_{22}] + 2\hat{\text{cov}}[\hat{\beta}_{02}, \hat{\beta}_{42}] + 2\hat{\text{cov}}[\hat{\beta}_{02}, \hat{\beta}_{52}] + \\ &2\hat{\text{cov}}[\hat{\beta}_{22}, \hat{\beta}_{42}] + 2\hat{\text{cov}}[\hat{\beta}_{22}, \hat{\beta}_{52}] + 2\hat{\text{cov}}[\hat{\beta}_{42}, \hat{\beta}_{52}] = 0,098. \end{aligned}$$

iii) O intervalo de 95% de confiança deste logito é, consequentemente,

$$-2,570 \pm 1,96 \times \sqrt{0,098} = (-3,183; -1,957).$$

iv) Analogamente, para o primeiro logito do modelo tem-se $\text{logito}[\hat{\pi}_{i11}] = 0,462$ e o intervalo de 95% de confiança (0,195; 0,729).

v) Finalmente, aplicando a expressão (2.11) com os limites do intervalo de confiança do logito, chega-se ao intervalo para a probabilidade predita:

• Limite inferior:

$$\frac{\exp(-3,183)}{1 + \exp(-3,183) + \exp(0,195)} = 0,0184.$$

• Limite superior:

$$\frac{\exp(-1,957)}{1 + \exp(-1,957) + \exp(0,729)} = 0,0440.$$

Portanto, a probabilidade de ocorrência de touceira nas condições determinadas é estimada em 0,0287, com um intervalo de 95% de confiança dado por (0,0184; 0,0440). Na Figura 4.5 são mostrados os intervalos de 95% de confiança para todas as probabilidades previstas pelo modelo. Ressalta-se aqui que, em um modelo ajustado sem que a dependência entre as observações fosse levada em consideração, estes intervalos de confiança seriam mais amplos pois os erros padrões das estimativas dos parâmetros seriam maiores.

Figura 4.5 – Intervalos de 95% de confiança para as probabilidades previstas pelo modelo 6, ajustado via EEG

O modelo ajustado proporciona várias comparações por meio de seus parâmetros estimados, sendo de grande valia o uso de razões de chances para comparar os diferentes níveis das variáveis selecionadas, auxiliando inclusive a evidenciar as interpretações feitas anteriormente. No cálculo das razões de chances e seus intervalos de confiança, entretanto, atenção especial deve ser dada às comparações envolvendo o fator pré-pastejo e a estação do ano devido à presença de interação entre eles.

Por exemplo, a razão de chances do pós-pastejo entre 45 cm *versus* 35 cm, comparando touceira e espaço vazio é dada simplesmente por

$$\hat{\psi} = \frac{\exp(\hat{\beta}_{01} + \hat{\beta}_{51})}{\exp(\hat{\beta}_{01})} = \exp(\hat{\beta}_{51})$$

com um intervalo de 95% confiança dado por $\left(\exp\left(\hat{\beta}_{51} \pm 1,96\sqrt{\hat{\text{Var}}[\beta_{51}]}\right) \right)$.

No entanto, a razão de chances do pré-pastejo entre 100% *versus* 95%, comparando touceira e espaço vazio, depende da estação do ano. Para o outono tem-se

$$\hat{\psi} = \frac{\exp(\hat{\beta}_{01} + \hat{\beta}_{41} + \hat{\beta}_{61} + \hat{\beta}_{11,1})}{\exp(\hat{\beta}_{01} + \hat{\beta}_{61})} = \exp(\hat{\beta}_{41} + \hat{\beta}_{11,1}),$$

de onde nota-se que $\ln(\hat{\psi}) = \hat{\beta}_{41} + \hat{\beta}_{11,1}$. A variância do logaritmo da razão de chances é estimada por $\hat{\text{Var}}[\ln(\hat{\psi})] = \hat{\text{Var}}[\hat{\beta}_{41}] + \hat{\text{Var}}[\hat{\beta}_{11,1}] + 2\text{cov}[\hat{\beta}_{41}, \hat{\beta}_{11,1}]$. Portanto, o intervalo de 95% de confiança para a razão de chances é dado por

$$\exp\left[\left(\hat{\beta}_{41} + \hat{\beta}_{11,1}\right) \pm 1,96\sqrt{\hat{\text{Var}}[\hat{\beta}_{41}] + \hat{\text{Var}}[\hat{\beta}_{11,1}] + 2\text{cov}[\hat{\beta}_{41}, \hat{\beta}_{11,1}]}\right].$$

Das diversas comparações possíveis (entre blocos, entre níveis de pré/pós-pastejo e entre estações do ano), neste texto é dada maior ênfase às interpretações que têm mais sentido prático, isto é, aquelas envolvendo o pré e o pós-pastejo, apresentadas na Tabela 4.6. A chance de ocorrer touceira ao invés de espaço vazio onde houve 100% de absorção de luminosidade é, **em média**, 1,10 (10% maior) a chance onde houve 95% de absorção para o primeiro verão, apesar de o intervalo de confiança conter o número 1 e portanto essa diferença não é significativa. Nas estações seguintes ocorre o contrário pois a chance de haver espaço vazio foi maior que a chance de haver touceira, comparando os mesmos níveis de absorção. Contudo, no outono e no inverno os intervalos de confiança também contêm o valor 1, não indicando diferença significativa; do início da primavera até o final do período analisado houve diferença significativa, apontando que 95% de absorção de luminosidade favorece o crescimento de touceiras ao invés de espaços vazios.

Tabela 4.6 – Estimativas das razões de chances ($\hat{\psi}$) entre os níveis das variáveis do modelo 6 ajustado via EEG, e respectivos intervalos de confiança (IC)

(a) Razões de chances envolvendo o fator pré-pastejo

Pré-pastejo (100% x 95%):						
Estação	Touceira x vazio		Invasora x vazio		Touceira x invasora	
	$\hat{\psi}$	IC _{95%} (ψ)	$\hat{\psi}$	IC _{95%} (ψ)	$\hat{\psi}$	IC _{95%} (ψ)
Verão 1	1,10	(0,79 ; 1,53)	0,75	(0,42 ; 1,34)	1,46	(0,84 ; 2,53)
Outono	0,77	(0,55 ; 1,07)	0,41	(0,22 ; 0,76)	1,85	(1,03 ; 3,34)
Inverno	0,79	(0,57 ; 1,09)	0,49	(0,24 ; 1,00)	1,59	(0,80 ; 3,18)
Início da primavera	0,61	(0,43 ; 0,85)	0,81	(0,36 ; 1,84)	0,74	(0,33 ; 1,66)
Fim da primavera	0,62	(0,45 ; 0,86)	0,92	(0,48 ; 1,73)	0,68	(0,36 ; 1,27)
Verão 2	0,51	(0,37 ; 0,71)	0,28	(0,14 ; 0,56)	1,83	(0,92 ; 3,63)

(b) Razões de chances envolvendo o fator pós-pastejo

Pós-pastejo (45 cm x 35 cm):					
Touceira x vazio		Invasora x vazio		Touceira x invasora	
$\hat{\psi}$	IC _{95%} (ψ)	$\hat{\psi}$	IC _{95%} (ψ)	$\hat{\psi}$	IC _{95%} (ψ)
1,05	(0,93 ; 1,20)	0,72	(0,54 ; 0,95)	1,46	(1,12 ; 1,91)

Seguindo a leitura da Tabela 4.6a nota-se que no outono e no segundo verão a chance de ocorrência de plantas invasoras foi significativamente superior à chance de ocorrência de espaços vazios nos locais onde houve 95% de absorção de luminosidade. E, por fim, no outono a chance de ocorrência de touceiras foi significativamente superior (em média 1,85 vez - 85% maior) à chance de ocorrência de plantas invasoras, onde houve 100% de absorção de luminosidade.

Fazendo um resumo por estação, comparando os níveis do fator pré-pastejo tem-se:

i) No verão 1 não há diferença significativa entre as chances de ocorrência dos 3 tipos de vegetação;

- ii) No outono, 100% de absorção favorece o aparecimento de espaços vazios ao invés de plantas invasoras. Nada se pode afirmar sobre as comparações dos demais tipos de vegetação;
- iii) No inverno não há diferença significativa entre as chances de ocorrência dos 3 tipos de vegetação;
- iv) Na primavera (início e fim), 95% de absorção favorece o aparecimento de touceiras ao invés de espaços vazios. Sobre as demais comparações desta estação, nada se pode afirmar.
- v) No verão 2, a chance de ocorrência de plantas invasoras ao invés de espaços vazios onde houve 100% de absorção foi (em média) 28% da chance de onde houve 95%, variando de 14% a 56% com 95% de confiança. Logo, 100% de absorção favorece a ocorrência de espaços vazios (ou 95% favorece a ocorrência de plantas invasoras).

Para os níveis de pós-pastejo as interpretações são mais simples, uma vez que não há interação com as estações do ano. Pela Tabela 4.6b observa-se que as diferenças significativas ocorreram quando se comparam as plantas invasoras em relação aos espaços vazios e as touceiras em relação às plantas invasoras. A chance de haver planta invasora ao invés de espaço vazio quando se utilizou 45 cm de altura do pasto é (em média) 72% da chance para 35 cm de altura; analogamente, a chance de ocorrência de touceiras ao invés de invasoras para 45 cm é 1,46 vez a chance para 35 cm, em média.

A interpretação prática destes resultados é que, para estimular o crescimento de touceiras em vez de plantas invasoras deve-se preferir 100% de absorção de luminosidade (apenas no outono) e/ou 45 cm de altura do pasto no pós-pastejo (em qualquer estação). Estas escolhas para o pré e pós-pastejo também favorecem o aparecimento de espaços vazios ao invés de plantas invasoras (no outono e segundo verão, para o pré-pastejo). Por último, se houver interesse em estimular o crescimento de touceiras ao invés de espaços vazios a sugestão é utilizar 95% de absorção de luminosidade na primavera e no segundo verão.

Por fim, e como descrito na metodologia, a dependência entre as observações feitas num mesmo ponto amostral é modelada por meio de razões de chances locais. Neste exemplo, em que a estrutura RC foi escolhida devido à grande variação entre os parâmetros intrínsecos, as estimativas dessas razões de chances locais são mostradas na matriz

0	0	1,20	0,79	1,42	0,66	1,00	1,00	1,01	0,99	2,89	0,43
0	0	0,79	1,36	0,66	1,64	1,00	1,00	0,99	1,01	0,43	1,93
1,20	0,79	0	0	3,13	0,42	2,24	0,57	2,60	0,30	2,78	0,32
0,79	1,36	0	0	0,42	1,92	0,57	1,47	0,30	4,54	0,32	3,61
1,42	0,66	3,13	0,42	0	0	2,64	0,31	1,13	0,76	1,00	1,00
0,66	1,64	0,42	1,92	0	0	0,31	4,04	0,76	1,87	1,00	1,04
1,00	1,00	2,24	0,57	2,64	0,31	0	0	1,83	0,54	3,87	0,29
1,00	1,00	0,57	1,47	0,31	4,04	0	0	0,54	1,85	0,29	3,10
1,01	0,99	2,60	0,30	1,13	0,76	1,83	0,54	0	0	1,10	0,88
0,99	1,01	0,30	4,54	0,76	1,87	0,54	1,85	0	0	0,88	1,18
2,89	0,43	2,78	0,32	1,00	1,00	3,87	0,29	1,10	0,88	0	0
0,43	1,93	0,32	3,61	1,00	1,04	0,29	3,10	0,88	1,18	0	0

A matriz é simétrica e a escolha pela estrutura RC implica que os blocos que a compõem são diferentes entre si. O segundo bloco na primeira linha relaciona o verão 1 e o outono, o terceiro bloco relaciona o verão 1 e o início da primavera, e assim sucessivamente. Para as interpretações desses resultados foram separados dois blocos da matriz, destacados em negrito.

Sejam as razões de chances locais dadas pelo bloco $\Psi_{12} = \begin{bmatrix} 1,20 & 0,79 \\ 0,79 & 1,36 \end{bmatrix}$:

- i) A chance de um ponto que era touceira no verão 1 vir a ser uma planta invasora ao invés de uma touceira no outono é a 1,20 (20% maior) que ocorrer o mesmo com um ponto que era planta invasora no verão 1.
- ii) A chance de um ponto que era touceira no verão 1 vir a ser um espaço vazio ao invés de uma planta invasora no outono é 0,79 (aproximadamente 20% menor) que ocorrer o mesmo com um ponto que era planta invasora no verão 1.
- iii) A chance de um ponto que era planta invasora no verão 1 vir a ser um espaço vazio no outono ao invés de continuar sendo invasora é 1,36 (36% maior) que ocorrer o mesmo com um ponto que era espaço vazio no verão 1.

Agora, para as razões de chances locais do bloco $\Psi_{34} = \begin{bmatrix} 2,64 & 0,31 \\ 0,31 & 4,04 \end{bmatrix}$:

- i) A chance de um ponto que era touceira no inverno vir a ser uma planta invasora ao invés de uma touceira no início da primavera é 2,64 vezes a chance de ocorrer o mesmo com um ponto que era planta invasora no inverno.
- ii) A chance de um ponto que era touceira no inverno vir a ser um espaço vazio ao invés de uma planta invasora no início da primavera é 31% da chance de ocorrer o mesmo com um ponto que era planta invasora no inverno.

iii) A chance de um ponto que era planta invasora no inverno vir a ser um espaço vazio no início da primavera ao invés de continuar sendo invasora é aproximadamente 4 vezes a chance de ocorrer o mesmo com um ponto que era espaço vazio no inverno.

É interessante notar, salvo algumas exceções, que as estimativas das razões de chances locais obtidas neste exemplo são bastante próximas de 1. Embora não haja no pacote `multgee` uma estrutura “independente” que possa ser testada, estes resultados permitem concluir que a dependência entre as observações de um mesmo indivíduo não é tão forte quanto se podia esperar. Mas, a fim de explorar um pouco mais os resultados do modelo, podem-se analisar os erros padrões *naive* e robustos das estimativas dos parâmetros. Como citado na seção 2.2.1.1, quando se leva em consideração a dependência entre as observações feitas em um mesmo indivíduo, os erros padrões tendem a diminuir e melhorar a precisão das estimativas (por isso são denominados robustos). Os erros padrões *naive* são obtidos quando não se leva em conta tal dependência. A Tabela 4.7 apresenta a comparação entre os dois tipos de erros padrões e nota-se que não há uma grande diferença entre eles, indicando que a associação entre as diferentes estações do ano é fraca. Porém, isso não significa que não seja necessário levar a dependência em consideração; apenas que ela é fraca e talvez não tão forte quanto se esperava no início do estudo. Um modelo ajustado sem levar em conta a dependência entre as estações do ano pode ser considerado errado.

Tabela 4.7 – Erros padrões (EP) *naive* e robustos das estimativas dos parâmetros referentes ao modelo 6, ajustado via EEG

(a) Valores associados ao primeiro logito do modelo

Parâmetro		Estimativa	EP <i>naive</i>	EP robusto	Diferença
Intercepto	β_{01}	0,265	0,140	0,137	-0,003
Bloco 2	β_{11}	0,174	0,101	0,092	-0,009
Bloco 3	β_{21}	0,050	0,098	0,089	-0,010
Bloco 4	β_{31}	0,118	0,101	0,096	-0,005
Pré (100%)	β_{41}	0,095	0,170	0,170	0,000
Pós (45 cm)	β_{51}	0,052	0,071	0,066	-0,005
Outono	β_{61}	0,259	0,173	0,174	0,000
Inverno	β_{71}	0,193	0,171	0,170	-0,001
Início da primavera	β_{81}	0,366	0,172	0,176	0,004
Fim da primavera	β_{91}	0,165	0,170	0,181	0,011
Verão 2	$\beta_{10,1}$	0,235	0,171	0,167	-0,005
Pré (100%) : Outono	$\beta_{11,1}$	-0,361	0,241	0,246	0,005
Pré (100%) : Inverno	$\beta_{12,1}$	-0,335	0,238	0,240	0,001
Pré (100%) : Início da primavera	$\beta_{13,1}$	-0,597	0,243	0,252	0,009
Pré (100%) : Fim da primavera	$\beta_{14,1}$	-0,576	0,239	0,246	0,008
Pré (100%) : Verão 2	$\beta_{15,1}$	-0,766	0,238	0,238	0,001

(b) Valores associados ao segundo logito do modelo

Parâmetro		Estimativa	EP <i>naive</i>	EP robusto	Diferença
Intercepto	β_{02}	-0,764	0,234	0,220	-0,014
Bloco 2	β_{12}	0,137	0,188	0,176	-0,013
Bloco 3	β_{22}	-1,195	0,244	0,233	-0,011
Bloco 4	β_{32}	-0,442	0,208	0,191	-0,016
Pré (100%)	β_{42}	-0,282	0,293	0,294	0,001
Pós (45 cm)	β_{52}	-0,329	0,151	0,142	-0,009
Outono	β_{62}	0,220	0,272	0,274	0,003
Inverno	β_{72}	-0,314	0,293	0,301	0,008
Início da primavera	β_{82}	-0,851	0,355	0,349	-0,006
Fim da primavera	β_{92}	-0,469	0,309	0,301	-0,007
Verão 2	$\beta_{10,2}$	-0,042	0,275	0,272	-0,003
Pré (100%) : Outono	$\beta_{11,2}$	-0,601	0,424	0,428	0,004
Pré (100%) : Inverno	$\beta_{12,2}$	-0,422	0,458	0,460	0,002
Pré (100%) : Início da primavera	$\beta_{13,2}$	0,075	0,510	0,516	0,006
Pré (100%) : Fim da primavera	$\beta_{14,2}$	0,194	0,437	0,444	0,006
Pré (100%) : Verão 2	$\beta_{15,2}$	-0,990	0,447	0,468	0,021

4.3 Modelo misto

4.3.1 Seleção e formulação do modelo

O modelo misto proposto é baseado no preditor linear selecionado no ajuste via EEG, ou seja, o modelo 6 da Tabela 3.2. Foram incluídos efeitos aleatórios somente no intercepto, por entender-se que não havia sentido prático em incluir mais efeitos aleatórios nas demais covariáveis. O modelo 6 é reescrito, portanto, como

$$\ln \left(\frac{\pi_{itk}}{\pi_{it3}} \right) = \left(\underbrace{\beta_{0k}}_{\text{intercepto}} + \underbrace{b_{ik}}_{\text{efeito aleatório}} \right) + \underbrace{\beta_{1k}X_{11i} + \beta_{2k}X_{12i} + \beta_{3k}X_{13i}}_{\text{bloco}} + \underbrace{\beta_{4k}X_{2i}}_{\text{pré}} + \underbrace{\beta_{5k}X_{3i}}_{\text{pós}} + \underbrace{\beta_{6k}X_{41t} + \beta_{7k}X_{42t} + \beta_{8k}X_{43t} + \beta_{9k}X_{44t} + \beta_{10,k}X_{45t}}_{\text{estação}} + \underbrace{\beta_{11,k}X_{2i}X_{41t} + \dots + \beta_{15,k}X_{2i}X_{45t}}_{\text{pré} \times \text{estação}}. \quad (4.2)$$

As estimativas dos parâmetros de efeitos fixos são apresentadas na Tabela 4.8, e as estimativas dos componentes de variância na Tabela 4.9. Nota-se que a estimativa da variância dos efeitos aleatórios do primeiro logito é extremamente baixa (praticamente nula), o que pode indicar que esse efeito não é significativo.

Como descrito nos métodos, ajustou-se também um modelo sem efeitos aleatórios para avaliar quão forte é a dependência entre as observações correlacionadas. O preditor linear deste modelo é idêntico à expressão (4.1), mas o ajuste é feito como um modelo de logitos generalizados convencional (seção 2.1.2.2), por máxima verossimilhança, uma vez que não há uma estrutura de associação para as observações. Apresentam-se aqui apenas alguns resultados deste modelo, uma vez que o foco são os modelos mistos (e marginais). A Tabela 4.10 contém algumas estatísticas e critérios do ajuste dos modelos misto e fixo, na qual se observa que há uma pequena diferença entre os dois modelos, com poucas alterações no AIC e na verossimilhança. Fazendo-se um teste de razão de verossimilhanças para avaliar se os efeitos aleatórios são significativos o p -valor equivalente é 0,1157, indicando que não o são. Desta forma, pode-se concluir que a dependência entre as observações de um mesmo ponto amostral é fraca (isso foi constatado também no ajuste via EEG, pelas razões de chances locais próximas a 1). Porém, como o objetivo é explorar o modelo misto, os resultados mostrados aqui são baseados no modelo (4.2).

Tabela 4.8 – Estimativas dos parâmetros e erros padrões (entre parênteses) do modelo misto ajustado, com a inclusão de efeitos aleatórios no modelo 6

(a) Primeiro logito - touceira × espaço vazio		
Parâmetro		Estimativa
Intercepto	β_{01}	0,261 (0,140)
Bloco 2	β_{11}	0,181 (0,099)
Bloco 3	β_{21}	0,055 (0,096)
Bloco 4	β_{31}	0,127 (0,099)
Pré (100%)	β_{41}	0,095 (0,170)
Pós (45 cm)	β_{51}	0,049 (0,069)
Outono	β_{61}	0,260 (0,174)
Inverno	β_{71}	0,193 (0,171)
Início da primavera	β_{81}^*	0,366 (0,172)
Fim da primavera	β_{91}	0,165 (0,170)
Verão 2	$\beta_{10,1}$	0,236 (0,172)
Pré (100%) : Outono	$\beta_{11,1}$	-0,361 (0,242)
Pré (100%) : Inverno	$\beta_{12,1}$	-0,335 (0,239)
Pré (100%) : Início da primavera	$\beta_{13,1}^*$	-0,596 (0,243)
Pré (100%) : Fim da primavera	$\beta_{14,1}^*$	-0,576 (0,239)
Pré (100%) : Verão 2	$\beta_{15,1}^*$	-0,767 (0,239)

(b) Segundo logito - planta invasora × espaço vazio		
Parâmetro		Estimativa
Intercepto	β_{02}^*	-0,853 (0,243)
Bloco 2	β_{12}	0,125 (0,182)
Bloco 3	β_{22}^*	-1,232 (0,239)
Bloco 4	β_{32}^*	-0,471 (0,201)
Pré (100%)	β_{42}	-0,280 (0,301)
Pós (45 cm)	β_{52}^*	-0,356 (0,147)
Outono	β_{62}	0,228 (0,280)
Inverno	β_{72}	-0,318 (0,304)
Início da primavera	β_{82}^*	-0,880 (0,359)
Fim da primavera	β_{92}	-0,486 (0,313)
Verão 2	$\beta_{10,2}$	-0,035 (0,290)
Pré (100%) : Outono	$\beta_{11,2}$	-0,641 (0,436)
Pré (100%) : Inverno	$\beta_{12,2}$	-0,449 (0,472)
Pré (100%) : Início da primavera	$\beta_{13,2}$	0,083 (0,515)
Pré (100%) : Fim da primavera	$\beta_{14,2}$	0,209 (0,443)
Pré (100%) : Verão 2	$\beta_{15,2}^*$	-1,021 (0,468)

Tabela 4.9 – Estimativas dos componentes de variância do modelo misto ajustado, com a inclusão de efeitos aleatórios no modelo 6

Parâmetro	Estimativa	Erro padrão
σ_1^2	$1,72 \times 10^{-18}$	-
σ_2^2	0,3112	0,1679
σ_{12}	0,0282	

Tabela 4.10 – Estatísticas de ajuste do modelo misto com a inclusão de efeitos aleatórios no modelo 6 e de um modelo sem efeitos aleatórios

Critério	Modelo misto	Modelo fixo
AIC	6.618,65	6.620,96
Verossimilhança (<i>l</i>)	-3.276,33	-3.278,48
Parâmetros	34	32

4.3.2 Qualidade do ajuste

Na Figura 4.6 são mostrados os valores observados e as curvas médias previstas pelo modelo, tomando-se a média dos valores preditos para cada ponto amostral. O ajuste do modelo é muito semelhante àquele obtido no ajuste via EEG, com as curvas muito próximas aos valores observados. Em seguida, na Figura 4.7, apresentam-se as correlações entre os valores estimados e observados, novamente muito similares às correlações obtidas pelo ajuste via EEG.

Figura 4.6 – Comparação das proporções observadas e previstas pelo modelo misto ajustado, com a inclusão de efeitos aleatórios no modelo 6

Figura 4.7 – Correlação entre os valores observados e preditos pelo modelo misto ajustado, com a inclusão de efeitos aleatórios no modelo 6

4.3.3 Interpretações e discussão

No modelo misto, cada ponto amostral tem sua própria probabilidade predita para cada tipo de vegetação. Na Figura 4.8, as linhas cinza representam as probabilidades individuais e as linhas mais grossas e coloridas representam o efeito médio de todos os pontos. Intervalos de confiança para as probabilidades médias preditas mostradas na Figura 4.6 são obtidos tomando-se a média dos intervalos de cada ponto amostral, e são mostrados na Figura 4.9.

Figura 4.8 – Probabilidades individuais e médias estimadas, considerando a presença de efeitos aleatórios ao predictor linear do modelo 6

Figura 4.9 – Intervalos de 95% de confiança para as probabilidades preditas pelo modelo misto, com a inclusão de efeitos aleatórios no modelo 6

A interpretação dos parâmetros do modelo ajustado é feita utilizando-se as razões de chances entre os níveis das covariáveis selecionadas no modelo. Porém, por se tratar de um MLGM, tais interpretações devem ser feitas a nível individual e não mais a nível médio populacional. No modelo selecionado, a única variável que varia ao longo do tempo para um

mesmo ponto são as estações do ano, sendo portanto a única que pode ser interpretada; ainda, devido à interação com o pré-pastejo, as interpretações devem levar isso em consideração. Como exemplo, o efeito do outono em relação ao verão 1 no primeiro logito (considerando 95% no pré-pastejo) é dado pela razão de chances

$$\hat{\psi} = \frac{\exp(\hat{\beta}_{01} + \hat{b}_{1i} + \hat{\beta}_{61})}{\exp(\hat{\beta}_{01} + \hat{b}_{1i})} = \exp(\hat{\beta}_{61}),$$

enquanto que esse efeito para 100% de absorção no pré-pastejo é dado por

$$\hat{\psi} = \frac{\exp(\hat{\beta}_{01} + \hat{b}_{1i} + \hat{\beta}_{41} + \hat{\beta}_{61} + \hat{\beta}_{11,1})}{\exp(\hat{\beta}_{01} + \hat{b}_{1i} + \hat{\beta}_{41})} = \exp(\hat{\beta}_{61} + \hat{\beta}_{11,1}).$$

As estimativas das razões de chances e respectivos intervalos de 95% de confiança são disponibilizados na Tabela 4.11. Devido à interação entre estação do ano e pré-pastejo fixou-se primeiramente o nível 95% de absorção de luminosidade e posteriormente 100%, e as razões de chances apresentadas referem-se à comparação de cada estação *versus* o verão 1. Podem ser obtidas as razões de chances comparando as estações duas a duas trabalhando-se com os exponenciais dos parâmetros, mas por haver muitas comparações possíveis optou-se por apresentar na Tabela 4.11 apenas as comparações em relação ao nível de referência (verão 1).

Tabela 4.11 – Estimativas das razões de chances ($\hat{\psi}$) entre as estações do ano de acordo com o modelo misto ajustado com base no modelo 6, com respectivos intervalos de confiança (IC)

(a) Razões de chances comparando as estações em relação ao primeiro verão, para o nível 95% no pré-pastejo

Pré-pastejo (95%):						
Comparação	Touceira x vazio		Invasora x vazio		Touceira x invasora	
	$\hat{\psi}$	IC _{95%} (ψ)	$\hat{\psi}$	IC _{95%} (ψ)	$\hat{\psi}$	IC _{95%} (ψ)
Outono × verão 1	1,30	(0,92 ; 1,82)	1,26	(0,73 ; 2,17)	1,03	(0,61 ; 1,74)
Inverno × verão 1	1,21	(0,87 ; 1,70)	0,73	(0,40 ; 1,32)	1,67	(0,94 ; 2,96)
Início primavera × verão 1	1,44	(1,03 ; 2,02)	0,41	(0,21 ; 0,84)	3,47	(1,76 ; 6,85)
Fim primavera × verão 1	1,18	(0,84 ; 1,65)	0,61	(0,33 ; 1,14)	1,92	(1,06 ; 3,47)
Verão 2 × verão 1	1,27	(0,90 ; 1,78)	0,97	(0,55 ; 1,71)	1,31	(0,76 ; 2,26)

(b) Razões de chances comparando as estações em relação ao primeiro verão, para o nível 100% no pré-pastejo

Pré-pastejo (100%):						
Comparação	Touceira x vazio		Invasora x vazio		Touceira x invasora	
	$\hat{\psi}$	IC _{95%} (ψ)	$\hat{\psi}$	IC _{95%} (ψ)	$\hat{\psi}$	IC _{95%} (ψ)
Outono × verão 1	0,90	(0,65 ; 1,26)	0,66	(0,48 ; 0,92)	0,90	(0,65 ; 1,25)
Inverno × verão 1	0,87	(0,63 ; 1,20)	0,46	(0,33 ; 0,64)	0,87	(0,63 ; 1,20)
Início primavera × verão 1	0,79	(0,57 ; 1,11)	0,45	(0,32 ; 0,63)	0,80	(0,57 ; 1,11)
Fim primavera × verão 1	0,66	(0,48 ; 0,92)	0,76	(0,55 ; 1,05)	0,66	(0,48 ; 0,92)
Verão 2 × verão 1	0,59	(0,43 ; 0,81)	0,35	(0,25 ; 0,48)	0,59	(0,43 ; 0,81)

Assim, a chance no início da primavera de em um ponto no campo ocorrer uma touceira ao invés de espaço vazio é 1,44 vez (44% maior) a chance no verão 1, para o nível de 95% de absorção de luminosidade - variando de 1,03 a 2,02 vezes com 95% de confiança. Comparando ainda essas duas estações, a chance de no início da primavera um ponto no campo ser uma planta invasora ao invés de espaço vazio é 41% da chance (menor, portanto) no verão 1.

Sobre uma razão de chances “média” para a população, sabe-se apenas que ela tem o mesmo “sentido” da razão de chances a nível individual, mas nada se pode afirmar sobre sua magnitude e intervalos de confiança. Assim, para as variáveis invariantes a nível individual, a única possibilidade para interpretar seus parâmetros é avaliar o sinal de suas estimativas para se ter uma ideia do comportamento médio dos indivíduos. Quando a estimativa de um parâmetro (ou de uma combinação de parâmetros) é positiva pode-se afirmar somente que a estimativa da razão de chances é maior que 1; quando for negativa, a razão de chances é menor que 1. A razão de chances “média” dos indivíduos apontará para o mesmo significado, mas sem uma precisão de quanto seja.

Logo, pode-se afirmar, por exemplo, que locais com 45 cm de altura do pasto no pós-pastejo têm chance média maior de desenvolver touceiras ao invés de espaço vazio do que locais com 35 cm de altura, pois $e^{\hat{\beta}_{51}} > 1$. Para o pré-pastejo no outono tem-se que $e^{(\hat{\beta}_{41} + \hat{\beta}_{11,1})} < 1$, o que significa que, em média, locais com 95% no pré-pastejo favorecem a formação de touceiras. Ainda é possível observar que os sinais das estimativas dos parâmetros do MLGM são os mesmos daqueles obtidos no ajuste via EEG; assim, conclui-se que o modelo leva às mesmas conclusões sobre quais tratamentos favorecem determinado tipo de vegetação.

5 CONSIDERAÇÕES FINAIS

O objetivo central deste trabalho é apresentar uma revisão dos principais modelos estatísticos disponíveis para a análise de dados politômicos nominais em estudos conduzidos de forma longitudinal, isto é, com várias observações tomadas ao longo do tempo nas unidades amostrais (indivíduos). As metodologias apresentadas concentram-se nos modelos marginais utilizando as equações de estimação generalizadas (EEG) e nos modelos que consideram efeitos aleatórios em sua composição, denominados modelos mistos. Realizou-se um paralelo entre as duas abordagens focando nas principais diferenças metodológicas e práticas que influenciam na escolha para cada situação.

Para o uso das EEG utilizou-se uma extensão atual (TOULOU MIS; AGRESTI; KATERI, 2013) da abordagem originalmente proposta por Liang e Zeger (1986) e mais conhecida na literatura, baseada no uso de razões de chances locais para descrever a dependência entre as observações de uma mesma unidade amostral. Quanto aos modelos mistos, embora a metodologia geral já seja bem fundamentada e conhecida na literatura, o seu emprego no contexto de dados multinomiais não é usual, assim como as interpretações decorrentes.

As duas metodologias foram aplicadas a um conjunto de dados proveniente de um experimento da área agrícola realizado em campo, sendo que ambas apresentaram resultados satisfatórios. Porém, devido à estrutura do experimento, bem como à natureza dos dados e relação de dependência fraca, as EEG foram mais bem exploradas do que o MLGM. Por essa razão conclui-se que, para esses dados, o modelo marginal ajustado via EEG é mais adequado por fornecer interpretações mais úteis. Em termos de qualidade do ajuste em relação aos valores observados em campo, ambos os modelos apresentaram resultados muito similares. Por fim, pode-se concluir sob as duas óticas (utilizando ferramentas distintas para cada uma) que a dependência entre as observações tomadas em um mesmo ponto do campo não é forte, mas não se pode ignorar a possibilidade de correlação em estudos longitudinais, o que reforça a importância destes métodos.

Contudo, na avaliação da qualidade do ajuste, em geral a área de dados categorizados não apresenta muitos recursos quanto à análise de resíduos dos modelos, sendo um campo de estudos atualmente. Os procedimentos implementados nos *softwares* também não apresentam muitos resultados sobre resíduos, o que caracteriza um potencial ponto de melhoria para estes modelos. No ajuste do MLGM pelo PROC GLIMMIX do SAS, por exemplo, não há nenhum tipo de resíduo disponível.

Notou-se ainda a carência de recursos computacionais nos *softwares* mais utilizados na área estatística, R e SAS, para a análise de dados com a referida estrutura. No R há apenas um pacote disponível para a análise via EEG, mas esta metodologia não se encontra (até o presente

momento) implementada no SAS. Para o ajuste via modelos mistos há dois pacotes no R, `mlogit` e `mixcat`, mas que apresentaram limitações e não foram utilizados neste trabalho, razão pela qual utilizou-se o PROC GLIMMIX do SAS.

Consideram-se ainda os intervalos de confiança obtidos para as probabilidades preditas e razões de chances como uma contribuição deste trabalho, uma vez que não foram encontrados na literatura muitos exemplos de como obtê-los, a não ser por exemplos simples e/ou para modelos com variáveis dicotômicas. O uso da metodologia atual para o modelo EEG também pode ser considerado uma contribuição, por não haver muitos exemplos de sua aplicação na literatura.

De forma geral, considera-se que os objetivos do trabalho foram atingidos e espera-se que este texto seja uma referência para o estudo de respostas politômicas nominais em estudos longitudinais.

REFERÊNCIAS

- AGRANONIK, M. **Equações de Estimação Generalizadas (GEE):** Aplicação em estudo sobre mortalidade neonatal em gemelares de Porto Alegre, RS (1995-2007). 2009. 110 p. Dissertação (Mestrado em Epidemiologia) – Universidade Federal do Rio Grande do Sul, 2009.
- AGRESTI, A. **Categorical Data Analysis**. 2nd. ed. New Jersey: John Wiley and Sons, 2002. 710 p.
- AGRESTI, A. **An Introduction to Categorical Data Analysis**. 2nd. ed. New Jersey: John Wiley and Sons, 2007. 372 p.
- AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, Tokyo, v. 19, n. 6, p. 716-723, Dec. 1974.
- BECKER, M. P.; CLOGG, C. C. Analysis of sets of 2-way contingency-tables using association models. **Journal of the American Statistical Association**, Gainesville, v. 84, n. 405, p. 142-151, Mar. 1989.
- BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. **Journal of the American Statistical Association**, Seattle, v. 88, n. 421, p. 9-25, Mar. 1993.
- CAREY, V. J.; LUMLEY, T.; RIPLEY, B. **gee: Generalized Estimation Equation solver**. 2012. Disponível em: <<http://CRAN.R-project.org/package=gee>>. Acesso em: 27 nov. 2015.
- CROISSANT, Y. **mlogit: multinomial logit model**. 2013. R package version 0.2-4. Disponível em: <<http://CRAN.R-project.org/package=mlogit>>. Acesso em: 27 nov. 2015.
- DEMÉTRIO, C. G. B.; CORDEIRO, G. M.; MORAL, R. A. **Modelos lineares generalizados e extensões**. Piracicaba: Setor de Ciências Exatas, 2014. 298 p.
- DIGGLE, P. J.; HEAGERTY, P.; LIANG, K. Y.; ZEGER, S. L. **Analysis of Longitudinal Data**. 2nd. ed. New York: Oxford University Press, 2002. 379 p.
- FREITAS, S.M.; ARAUJO, A.M.S.; LARA, I.A.R. R procedures for multinomial categorical data: Discussion and suggestions. In: INTERNATIONAL BIOMETRIC CONFERENCE, 27., 2014, Florence. **Abstracts...** Florence: International Biometric Society, 2014. p. 6-11.
- GIOLO, S. R. **Introdução à análise de dados categóricos com aplicações**. Curitiba: Departamento de Estatística, 2012. 190 p. Disponível em: <www.ufpr.br/~giolo>. Acesso em: 27 nov. 2015.
- GOODMAN, L. A. The analysis of cross-classified data having ordered and or unordered categories - association models, correlation models, and asymmetry models for contingency-tables with or without missing entries. **Annals of Statistics**, Chicago, v. 13, n. 1, p. 10-69, Mar. 1985.
- HARTZEL, J.; AGRESTI, A.; CAFFO, B. Multinomial logit random effects models. **Statistical Modelling**, West Point, v. 1, p. 81-102, Jul. 2001.

- HEAGERTY, P. J.; ZEGER, S. L. Marginal regression models for clustered ordinal measurements. **Journal of the American Statistical Association**, Seattle, v. 91, n. 435, p. 1024-1036, 1996.
- HEDEKER, D. A mixed-effects multinomial logistic regression model. **Statistics in Medicine**, Chicago, v. 22, n. 9, p. 1433-1446, 2003.
- HEDEKER, D.; GIBBONS, R. D. A random-effects ordinal regression-model for multilevel analysis. **Biometrics**, Chicago, v. 50, n. 4, p. 933-944, 1994.
- HØJSGAARD, S.; HALEKOH, U.; YAN, J. The R package geepack for generalized estimating equations. **Journal of Statistical Software**, Tjele, v. 15, n. 2, p. 1-11, 2006.
- LAIRD, N. M.; WARE, J. H. Random-Effects Models for Longitudinal Data. **Biometrics**, Boston, v. 38, n. 4, p. 963-974, Dec. 1982.
- LARA, I. A. R. **Modelos de transição para dados binários**. 2007. 128 p. Tese (Doutorado em Agronomia) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, 2007.
- LIANG, K. Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. **Biometrika**, Baltimore, v. 73, n. 1, p. 13-22, Apr. 1986.
- LIPSITZ, S. R.; KIM, K.; ZHAO, L. P. Analysis of repeated categorical data using generalized estimating equations. **Statistics in Medicine**, Boston, v. 13, n. 11, p. 1149-1163, 1994.
- LOCASCIO, J. J.; ATRI, A. An overview of longitudinal data analysis methods for neurological research. **Dement Geriatr Cogn Disord Extra**, Boston, v. 1, p. 330-357, Oct. 2011.
- MCCULLAGH, P. Quasi-likelihood functions. **Annals of Statistics**, London, v. 11, n. 1, p. 59-67, 1983.
- MILLER, M. E.; DAVIS, C. S.; LANDIS, J. R. The analysis of longitudinal polytomous data - generalized estimating equations and connections with weighted least-squares. **Biometrics**, Winston-Salem, v. 49, n. 4, p. 1033-1044, 1993.
- MOLENBERGHS, G.; VERBEKE, G. **Models for Discrete Longitudinal Data**. New York: Springer, 2005. 683 p.
- NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society Series A-General**, Harpenden, v. 135, n. 3, p. 370-384, 1972.
- NEUHAUS, J. M.; KALBFLEISCH, J. D.; HAUCK, W. W. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. **International Statistical Review**, San Francisco, v. 59, n. 1, p. 25-35, Apr. 1991.
- PAN, W. Akaike's information criterion in generalized estimating equations. **Biometrics**, Minneapolis, v. 57, n. 1, p. 120-125, Mar. 2001.
- PAPAGEORGIOU, G.; HINDE, J. **mixcat**: Mixed effects cumulative link and logistic regression models. 2012. R package version 1.0-3. Disponível em: <<http://CRAN.R-project.org/package=mixcat>>. Acesso em: 27 nov. 2015.
- PAPAGEORGIOU, G.; HINDE, J. Multivariate generalized linear mixed models with semi-nonparametric and smooth nonparametric random effects densities. **Statistics and Computing**, Galway, v. 22, n. 1, p. 79-92, 2012.

- PARZEN, M.; GHOSH, S.; LIPSITZ, S.; SINHA, D.; FITZMAURICE, G. M.; MALLICK, B. K.; IBRAHIM, J. G. A generalized linear mixed model for longitudinal binary data with a marginal logit link function. **Annals of Applied Statistics**, Texas, v. 5, n. 1, p. 449-467, Mar. 2011.
- PEREIRA, L.E.T; PAIVA, A.J.; GEREMIA, E.V.; SILVA, S.C. Grazing management and tussock distribution in elephant grass. **Grass and Forage Science**, Wiley Online Library, 2014.
- PINHEIRO, J. C.; BATES, D. M. **Mixed-Effects Models in S and S-Plus**. New York: Springer-Verlag, 2000. 528 p.
- R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2014. Disponível em: <<http://www.R-project.org/>>. Acesso em: 27 nov. 2015.
- SANTANA, S. P. B. **Modelos estatísticos para análise de dados longitudinais categorizados ordinais**. 2004. 86 p. Dissertação (Mestrado em Estatística) – Universidade Federal de Pernambuco, 2004.
- SAS Institute Inc. **SAS/STAT User's Guide, version 9.3**. Cary NC, 2011.
- SAS Institute Inc. **The GENMOD Procedure**. Cary NC, 2008a.
- SAS Institute Inc. **The GLIMMIX Procedure**. Cary NC, 2008b.
- SILVA, N. B. da. **Diferentes estratégias para modelagem de respostas politômicas ordinais em estudos longitudinais**. 2013. 66 p. Dissertação (Mestrado em Estatística) – Universidade Federal de Minas Gerais, 2013.
- TOULOUMIS, A. **Generalized Estimating Equations for Multinomial Responses**. 2011. 196 p. Dissertation (Doctorship in Philosophy) – University of Florida, 2011. Disponível em: <<http://ufdc.ufl.edu/UFE0043226/00001>>. Acesso em: 23 nov. 2015.
- TOULOUMIS, Anestis. R package multgee: A Generalized Estimating Equations Solver for Multinomial Responses. **Journal of Statistical Software**, Cambridge, v. 64, n. 8, p. 1-14, Mar. 2015.
- TOULOUMIS, A.; AGRESTI, A.; KATERI, M. GEE for Multinomial Responses Using a Local Odds Ratios Parameterization. **Biometrics**, Hinxton, v. 69, n. 3, p. 633-640, Sep. 2013.
- TWISK, J. **Applied Longitudinal Data Analysis for Epidemiology**. Cambridge: Cambridge University Press, 2003. 301 p.
- VERBEKE, G.; FIEUWS, S.; MOLENBERGHS, G.; DAVIDIAN, M. The analysis of multivariate longitudinal data: A review. **Statistical Methods in Medical Research**, Leuven, v. 23, n. 1, p. 42-59, 2014.
- VERBEKE, G.; MOLENBERGHS, G. **Linear Mixed Models for Longitudinal Data**. New York: Springer-Verlag, 2000. 568 p.
- VERBEKE, G.; MOLENBERGHS, G. WORKSHOP ON LONGITUDINAL AND INCOMPLETE DATA. 2014, Piracicaba. **Introduction to longitudinal data analysis**. Leuven: Katholieke Universiteit Leuven and Universiteit Hasselt, 2014. 336 p.

VERBEKE, G.; MOLENBERGHS, G.; RIZOPOULOS, D. Models for Longitudinal Data. In: K. van Monrfort et al. **Longitudinal research with latent variables**. Leuven: Springer, 2010. chap. 2, p. 37-96.

WEDDERBURN, R. W. M. Quasi-likelihood functions, generalized linear models, and Gauss-Newton method. **Biometrika**, Harpenden, v. 61, n. 3, p. 439-447, 1974.

WILLIAMSON, J. M.; LIPSITZ, S. R.; KIM, K. M. GEECAT and GEEGOR: computer programs for the analysis of correlated categorical response data. **Computer Methods and Programs in Biomedicine**, Atlanta, v. 58, n. 1, p. 25-34, 1999.

ZEGER, S. L.; LIANG, K. Y. An overview of methods for the analysis of longitudinal data. **Statistics in Medicine**, Baltimore, v. 11, n. 14-15, p. 1825-1839, 1992.

ZEGER, S. L.; LIANG, K. Y.; ALBERT, P. S. Models for longitudinal data - a generalized estimating equation approach. **Biometrics**, Baltimore, v. 44, n. 4, p. 1049-1060, Dec. 1988.

APÊNDICES

APÊNDICE A - PRINCIPAIS PACOTES DO *SOFTWARE R*

Aqui são apresentados os principais pacotes e funções do *software R* para a análise de dados categorizados politômicos (nominais e ordinais) e/ou longitudinais. Também é informado se o pacote possibilita a inclusão de efeitos aleatórios e se a estrutura dos dados suportada é individual ou agrupada.

Pacote	Função	Estudo	Resposta	Efeitos aleatórios	Estrutura dos dados
nnet	multinom()	Transversal	Nominal	Não	Individual Agrupada
VGAM	vglm()	Transversal	Nominal Ordinal	Não	Individual Agrupada
mlogit	mlogit()	Transversal Longitudinal	Nominal Ordinal	Sim	Individual
multinomRob	multinomMLE()	Transversal	Nominal	Não	Individual Agrupada
ordinal	clm()	Transversal	Ordinal	Não	Individual
ordinal	clmm2()	Transversal Longitudinal	Ordinal	Sim	Individual
gee/geepack	geeglm()	Longitudinal	Família MLG	Não	Individual
multgee	nomLORgee()	Longitudinal	Nominal	Não	Individual
multgee	ordLORgee()	Longitudinal	Ordinal	Não	Individual
drm	drm()	Longitudinal	Nominal Ordinal	Não	Individual
MASS	glmmPQL()	Transversal Longitudinal	Família MLG	Sim	Individual
repeated	glmm()	Transversal Longitudinal	Família MLG	Sim	Individual
lme4	glmer()	Transversal Longitudinal	Família MLG	Sim	Individual Agrupada
correlatedData	geeDesign()	Longitudinal	Nominal Ordinal	Não	Individual

Fonte: adaptação de Freitas, Araujo e Lara (2014).

ANEXOS

ANEXO A: Comandos para o ajuste do modelo via EEG - *software R*

```
library(ggplot2) # pacote utilizado para os gráficos
library(doby) # funções úteis para a análise exploratória
library(multgee) # ajuste do modelo segundo a abordagem de EEG

## 1) Leitura dos dados:
# Dados na visão 1 (há uma variável 'tipo' que identifica o tipo de vegetação):
dados = read.table('Dados_visao1.txt',header=TRUE,dec=',',sep='\t')
dados$estacao = factor(dados$estacao, levels=c('Verão 1','Outono','Inverno',
+ 'Primavera - início','Primavera - fim','Verão 2'))
dados$tipo = factor(dados$tipo, levels=c('Planta invasora','Espaço vazio','Touceira'))
dados$pre = as.factor(dados$pre)
dados$pos = as.factor(dados$pos)
dados$bloco = as.factor(dados$bloco)

# Dados na visão 2 (há 3 colunas com as contagens de quantos pontos foram classificados
em cada tipo de vegetação):
dados2 = read.table('Dados_visao2.txt',header=TRUE,dec=',',sep='\t')
dados2$estacao = factor(dados2$estacao, levels=c('Verão 1','Outono','Inverno',
+ 'Primavera - início','Primavera - fim','Verão 2'))
dados2$pre = as.factor(dados2$pre)
dados2$pos = as.factor(dados2$pos)
dados2$bloco = as.factor(dados2$bloco)

## 2) Análise descritiva:
# Proporções médias e desvios padrões (tabela 4.3):
summaryBy((touceira/total) ~ cod_estacao + combin, data=dados2,
+ FUN = c(mean,sd), na.rm=TRUE)
summaryBy((invasora/total) ~ cod_estacao + combin, data=dados2,
+ FUN = c(mean,sd), na.rm=TRUE)
summaryBy((vazio/total) ~ cod_estacao + combin, data=dados2,
+ FUN = c(mean,sd), na.rm=TRUE)

# Construção do gráfico da figura 4.1:
prop_touceira = as.data.frame(summaryBy((touceira/total) ~ pre + pos + cod_estacao,
+ data=dados2, FUN = c(mean), na.rm=TRUE))
prop_invasora = as.data.frame(summaryBy((invasora/total) ~ pre + pos + cod_estacao,
+ data=dados2, FUN = c(mean), na.rm=TRUE))
prop_vazio = as.data.frame(summaryBy((vazio/total) ~ pre + pos + cod_estacao,
+ data=dados2, FUN = c(mean), na.rm=TRUE))

aux = data.frame(pre = prop_touceira[,1], pos = prop_touceira[,2],
estacao = prop_touceira[,3], tipo = rep(c('Touceira','Planta invasora','Espaço vazio'),
+ each=length(prop_touceira[,1])),prop=c(prop_touceira[,4],prop_invasora[,4],
```



```

+ prop_vazio[,4]))
levels(aux$pre)[levels(aux$pre)=='95'] = 'Pré-pastejo: 95%'
levels(aux$pre)[levels(aux$pre)=='100'] = 'Pré-pastejo: 100%'
levels(aux$pos)[levels(aux$pos)=='35'] = 'Pós-pastejo: 35 cm'
levels(aux$pos)[levels(aux$pos)=='45'] = 'Pós-pastejo: 45 cm'

ggplot(aux, aes(x=estacao,y=prop,colour=tipo,group=tipo)) + geom_line(size=1)
+ facet_wrap(~pre+pos,ncol=2) + geom_point(size=4,shape=21,fill='white')
+ xlab('\nEstação') + ylab('Proporção média\n')

## 3) Modelagem via EEG (pacote MULTGEE):
# 3.1) Escolha da estrutura de associação para as razões de chances locais:
newdata = with(dados,data.frame(y=cod_tipo, id=ponto_experimento, repeated=cod_estacao))
intrinsic.pars(y=y, data=newdata, id=id, repeated=repeated, rscale='nominal')

# 3.2) Escolha das variáveis (preditor linear):
# Começando com a estrutura de um fatorial mais o tempo, com interação tripla:
mod_gee1 = nomLORgee(cod_tipo ~ bloco + pre*pos*cod_estacao, data=dados, id=ponto_experimento,
+ repeated=cod_estacao, LORstr = 'RC')

# Testando sem a interação tripla:
mod_gee2 = update(mod_gee1, formula = ~. - pre:pos:cod_estacao)
waldts(mod_gee1,mod_gee2)

# Testando sem a interação dos fatores com o tempo, tirando as duas ao mesmo tempo:
mod_gee3 = update(mod_gee2, formula = ~. - pre:cod_estacao - pos:cod_estacao)
waldts(mod_gee3,mod_gee2)

# Testando sem uma das interações:
mod_gee4 = update(mod_gee2, formula = ~. - pre:cod_estacao)
waldts(mod_gee4,mod_gee2)

# Testando sem a outra interação:
mod_gee5 = update(mod_gee2, formula = ~. - pos:cod_estacao)
waldts(mod_gee5,mod_gee2)

# Testando se dá para retirar a interação pré x pós:
mod_gee6 = update(mod_gee4, formula = ~. - pre:pos)
waldts(mod_gee5,mod_gee6)

# Testando sem o pós-pastejo:
mod_gee7 = update(mod_gee6, formula = ~. - pos)
waldts(mod_gee6,mod_gee7)

mod_gee_final = mod_gee6
summary(mod_gee_final)

```

```

# 3.3) Comparando os erros padrões 'naive' e robustos:
ep_gee = cbind(as.data.frame(sqrt(diag(mod_gee_final$naive.variance))),
+ as.data.frame(sqrt(diag(mod_gee_final$robust.variance))))
colnames(ep_gee) = c('naive', 'robust')
ep_gee$dif = ep_gee[,2] - ep_gee[,1]
ep_gee

# 3.4) Valores preditos:
preditos = as.data.frame(mod_gee_final$fitted.values)
str(preditos)
names(preditos) = c('p_touceira', 'p_invasora', 'p_vazio')

# 3.5) Resíduos:
residuos = as.data.frame(mod_gee_final$residuals)
str(residuos)

residuos$ponto = c(rep(1:480, each=6), rep(481:520, each=5), rep(521:640, each=6))
names(residuos)[1] = 'logito1'
names(residuos)[2] = 'logito2'

media_res1 = with(residuos, mean(logito1))
media_res2 = with(residuos, mean(logito2))

residuos2 = data.frame(ponto = residuos$ponto, residuo=c(residuos$logito1,
+ residuos$logito2), logito=rep(c('Logito 1', 'Logito 2'), each=length(residuos$ponto)))

medias_res = data.frame(logito=levels(residuos2$logito), media=c(media_res1, media_res2))

ggplot(residuos2, aes(x=ponto, y=residuo)) + geom_point(shape=21, size=3, fill=grey(0.6))
+ facet_wrap(~logito) + xlab('\nPonto amostral') + ylab('Resíduo\n') +
geom_hline(aes(yintercept=media), data=medias_res, color='red', size=1.3, linetype='dashed')

# 3.6) Intervalos de confiança (IC) para as probabilidades:

# Matriz de covariâncias das estimativas dos parâmetros:
vcov_gee = mod_gee_final$robust.variance
vcov_1.1 = vcov_gee[1:16, 1:16] # matriz do logito 1
vcov_2.1 = vcov_gee[17:32, 17:32] # matriz do logito 2

# Ajustando o modelo novamente, mas fixando 'touceira' como referência para poder pegar
o erro padrão das estimativas relacionadas à probabilidade de acontecer 'vazio':
mod_gee_final2 = nomLORgee(cod_tipo2 ~ bloco + pre + pos + cod_estacao
+ pre:cod_estacao, data=dados, id=ponto_experimento, repeated=cod_estacao, LORstr = 'RC')

```

```

summary(mod_gee_final2)

# Matriz de covariâncias das estimativas dos parâmetros (segundo modelo):
vcov_gee2 = mod_gee_final2$robust.variance
vcov_1.2 = vcov_gee2[1:16,1:16] # matriz do logito 1, mas do segundo modelo
vcov_2.2 = vcov_gee2[17:32,17:32] # matriz do logito 2, mas do segundo modelo

# Vetores de valores 0 e 1 criado pelo autor para auxiliar a construir os IC matricialmente,
conforme descrito nos métodos:
coef_aux = read.table('Coeficientes_auxiliares.txt',header=TRUE,dec=',',sep='\t')

betas_1.1 = as.matrix(coef(mod_gee_final)[1:16]) # estimativas dos parâmetros do logito 1
betas_2.1 = as.matrix(coef(mod_gee_final)[17:32]) # estimativas dos parâmetros do logito 2
betas_1.2 = as.matrix(coef(mod_gee_final2)[1:16]) # estimativas dos parâmetros do logito 1
do segundo modelo (vazio x touceira)
betas_2.2 = as.matrix(coef(mod_gee_final2)[17:32]) # estimativas dos parâmetros do logito 2
do segundo modelo (invasora x touceira)

crit = qnorm(1-0.05/2) #95 % de confiança
ic = c()

for (i in 2:96){
x = coef_aux[,i];x

l1.1 = t(x) %%% betas_1.1 # valor do logito 1
l2.1 = t(x) %%% betas_2.1 # valor do logito 2
l1.2 = t(x) %%% betas_1.2 # valor do logito 1, do modelo 2
l2.2 = t(x) %%% betas_2.2 # valor do logito 2, do modelo 2

ep_l1.1 = sqrt(t(x) %%% vcov_1.1 %%% x) # erro padrão do logito 1
ep_l2.1 = sqrt(t(x) %%% vcov_2.1 %%% x) # erro padrão do logito 2
ep_l1.2 = sqrt(t(x) %%% vcov_1.2 %%% x) # erro padrão do logito 1
ep_l2.2 = sqrt(t(x) %%% vcov_2.2 %%% x) # erro padrão do logito 2

li_l1.1 = l1.1 - crit*ep_l1.1 # limite inferior do logito 1
ls_l1.1 = l1.1 + crit*ep_l1.1 # limite superior do logito 1
li_l2.1 = l2.1 - crit*ep_l2.1 # limite inferior do logito 2
ls_l2.1 = l2.1 + crit*ep_l2.1 # limite superior do logito 2

li_l1.2 = l1.2 - crit*ep_l1.2 # limite inferior do logito 1, do modelo 2
ls_l1.2 = l1.2 + crit*ep_l1.2 # limite superior do logito 1, do modelo 2
li_l2.2 = l2.2 - crit*ep_l2.2 # limite inferior do logito 2, do modelo 2
ls_l2.2 = l2.2 + crit*ep_l2.2 # limite superior do logito 2, do modelo 2

# Limites inferior e superior da probabilidade de touceira:
li_p_touc = exp(li_l1.1)/(1 + exp(li_l1.1) + exp(li_l2.1))
ls_p_touc = exp(ls_l1.1)/(1 + exp(ls_l1.1) + exp(ls_l2.1))

```

```

# Limites inferior e superior da probabilidade de invasora:
li_p_inv = exp(li_l2.1)/(1 + exp(li_l1.1) + exp(li_l2.1))
ls_p_inv = exp(ls_l2.1)/(1 + exp(ls_l1.1) + exp(ls_l2.1))

# Limites inferior e superior da probabilidade de espaço vazio:
li_p_vazio = exp(li_l1.2)/(1 + exp(li_l1.2) + exp(li_l2.2))
ls_p_vazio = exp(ls_l1.2)/(1 + exp(ls_l1.2) + exp(ls_l2.2))

ic = rbind(ic,cbind(li_p_touc,ls_p_touc,li_p_inv,ls_p_inv,li_p_vazio,ls_p_vazio))
}
ic = as.data.frame(ic)
names(ic) = c('p_touceira_li','p_touceira_ls','p_invasora_li','p_invasora_ls',
+ 'p_vazio_li','p_vazio_ls')

# o resultado do objeto 'ic' é um data.frame com os intervalos de confiança de
cada tipo de vegetação para cada combinação de bloco, pré, pós-pastejo e estação do ano

# 3.7) Razões de chances (comparação dos níveis de pré e pós-pastejo):

# Pós-pastejo (45 x 35 cm): --> independe da estação pois não há interação
# Touceira x vazio (logito 1):
(razao_est = exp(betas_1.1[6]))
(ep = sqrt(vcov_1.1[6,6]))
(razao_li = exp(betas_1.1[6] - crit*ep))
(razao_ls = exp(betas_1.1[6] + crit*ep))
cbind(razao_li,razao_est,razao_ls)

# Invasora x vazio (logito 2):
(razao_est = exp(betas_2.1[6]))
(ep = sqrt(vcov_2.1[6,6]))
(razao_li = exp(betas_2.1[6] - crit*ep))
(razao_ls = exp(betas_2.1[6] + crit*ep))
cbind(razao_li,razao_est,razao_ls)

# Touceira x invasora (logito 2 do modelo 2, com o sinal dos coeficientes invertido):
(razao_est = exp((-1)*betas_2.2[6]))
(ep = sqrt(vcov_2.2[6,6]))
(razao_li = exp((-1)*betas_2.2[6] - crit*ep))
(razao_ls = exp((-1)*betas_2.2[6] + crit*ep))
cbind(razao_li,razao_est,razao_ls)

# Pré-pastejo (100% x 95%) --> depende da estação pois há interação.
# Touceira x vazio (logito 1):
razoes_1 = c()

```

```

# Verão 1:
(razao_est = exp(betas_1.1[5]))
(ep = sqrt(vcov_1.1[5,5]))
(razao_li = exp(betas_1.1[5] - crit*ep))
(razao_ls = exp(betas_1.1[5] + crit*ep))
razoes_1 = rbind(razoes_1,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

# Outono:
(razao_est = exp(betas_1.1[5] + betas_1.1[12]))
(ep = sqrt(vcov_1.1[5,5] + vcov_1.1[12,12] + 2*vcov_1.1[5,12]))
(razao_li = exp(betas_1.1[5] + betas_1.1[12] - crit*ep))
(razao_ls = exp(betas_1.1[5] + betas_1.1[12] + crit*ep))
razoes_1 = rbind(razoes_1,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

# Inverno:
(razao_est = exp(betas_1.1[5] + betas_1.1[13]))
(ep = sqrt(vcov_1.1[5,5] + vcov_1.1[13,13] + 2*vcov_1.1[5,13]))
(razao_li = exp(betas_1.1[5] + betas_1.1[13] - crit*ep))
(razao_ls = exp(betas_1.1[5] + betas_1.1[13] + crit*ep))
razoes_1 = rbind(razoes_1,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

# Início - primavera:
(razao_est = exp(betas_1.1[5] + betas_1.1[14]))
(ep = sqrt(vcov_1.1[5,5] + vcov_1.1[14,14] + 2*vcov_1.1[5,14]))
(razao_li = exp(betas_1.1[5] + betas_1.1[14] - crit*ep))
(razao_ls = exp(betas_1.1[5] + betas_1.1[14] + crit*ep))
razoes_1 = rbind(razoes_1,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

# Fim - primavera:
(razao_est = exp(betas_1.1[5] + betas_1.1[15]))
(ep = sqrt(vcov_1.1[5,5] + vcov_1.1[15,15] + 2*vcov_1.1[5,15]))
(razao_li = exp(betas_1.1[5] + betas_1.1[15] - crit*ep))
(razao_ls = exp(betas_1.1[5] + betas_1.1[15] + crit*ep))
razoes_1 = rbind(razoes_1,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

# Verão 2:
(razao_est = exp(betas_1.1[5] + betas_1.1[16]))
(ep = sqrt(vcov_1.1[5,5] + vcov_1.1[16,16] + 2*vcov_1.1[5,16]))
(razao_li = exp(betas_1.1[5] + betas_1.1[16] - crit*ep))
(razao_ls = exp(betas_1.1[5] + betas_1.1[16] + crit*ep))
razoes_1 = rbind(razoes_1,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

```

```

razoes_1

# Invasora x vazio (logito 2):
razoes_2 = c()
# Verão 1:
(razao_est = exp(betas_2.1[5]))
(ep = sqrt(vcov_2.1[5,5]))
(razao_li = exp(betas_2.1[5] - crit*ep))
(razao_ls = exp(betas_2.1[5] + crit*ep))
razoes_2 = rbind(razoes_2,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ;
+ ',round(razao_ls,2),')'))))

# Outono:
(razao_est = exp(betas_2.1[5] + betas_2.1[12]))
(ep = sqrt(vcov_2.1[5,5] + vcov_2.1[12,12] + 2*vcov_2.1[5,12]))
(razao_li = exp(betas_2.1[5] + betas_2.1[12] - crit*ep))
(razao_ls = exp(betas_2.1[5] + betas_2.1[12] + crit*ep))
razoes_2 = rbind(razoes_2,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ;
+ ',round(razao_ls,2),')'))))

# Inverno:
(razao_est = exp(betas_2.1[5] + betas_2.1[13]))
(ep = sqrt(vcov_2.1[5,5] + vcov_2.1[13,13] + 2*vcov_2.1[5,13]))
(razao_li = exp(betas_2.1[5] + betas_2.1[13] - crit*ep))
(razao_ls = exp(betas_2.1[5] + betas_2.1[13] + crit*ep))
razoes_2 = rbind(razoes_2,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ;
+ ',round(razao_ls,2),')'))))

# Início - primavera:
(razao_est = exp(betas_2.1[5] + betas_2.1[14]))
(ep = sqrt(vcov_2.1[5,5] + vcov_2.1[14,14] + 2*vcov_2.1[5,14]))
(razao_li = exp(betas_2.1[5] + betas_2.1[14] - crit*ep))
(razao_ls = exp(betas_2.1[5] + betas_2.1[14] + crit*ep))
razoes_2 = rbind(razoes_2,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ;
+ ',round(razao_ls,2),')'))))

# Fim - primavera:
(razao_est = exp(betas_2.1[5] + betas_2.1[15]))
(ep = sqrt(vcov_2.1[5,5] + vcov_2.1[15,15] + 2*vcov_2.1[5,15]))
(razao_li = exp(betas_2.1[5] + betas_2.1[15] - crit*ep))
(razao_ls = exp(betas_2.1[5] + betas_2.1[15] + crit*ep))
razoes_2 = rbind(razoes_2,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ;
+ ',round(razao_ls,2),')'))))

# Verão 2:
(razao_est = exp(betas_2.1[5] + betas_2.1[16]))

```

```

(ep = sqrt(vcov_2.1[5,5] + vcov_2.1[16,16] + 2*vcov_2.1[5,16]))
(razao_li = exp(betas_2.1[5] + betas_2.1[16] - crit*ep))
(razao_ls = exp(betas_2.1[5] + betas_2.1[16] + crit*ep))
razoes_2 = rbind(razoes_2,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

razoes_2

# Touceira x invasora (logito 2 do modelo 2, com os sinais dos coeficientes invertidos):
razoes_3 = c()
# Verão 1:
(razao_est = exp((-1)*betas_2.2[5]))
(ep = sqrt(vcov_2.2[5,5]))
(razao_li = exp((-1)*betas_2.2[5] - crit*ep))
(razao_ls = exp((-1)*betas_2.2[5] + crit*ep))
razoes_3 = rbind(razoes_3,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

# Outono:
(razao_est = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[12]))
(ep = sqrt(vcov_2.2[5,5] + vcov_2.2[12,12] + 2*vcov_2.2[5,12]))
(razao_li = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[12] - crit*ep))
(razao_ls = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[12] + crit*ep))
razoes_3 = rbind(razoes_3,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

# Inverno:
(razao_est = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[13]))
(ep = sqrt(vcov_2.2[5,5] + vcov_2.2[13,13] + 2*vcov_2.2[5,13]))
(razao_li = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[13] - crit*ep))
(razao_ls = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[13] + crit*ep))
razoes_3 = rbind(razoes_3,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

# Início - primavera:
(razao_est = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[14]))
(ep = sqrt(vcov_2.2[5,5] + vcov_2.2[14,14] + 2*vcov_2.2[5,14]))
(razao_li = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[14] - crit*ep))
(razao_ls = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[14] + crit*ep))
razoes_3 = rbind(razoes_3,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))

# Fim - primavera:
(razao_est = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[15]))
(ep = sqrt(vcov_2.2[5,5] + vcov_2.2[15,15] + 2*vcov_2.2[5,15]))
(razao_li = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[15] - crit*ep))
(razao_ls = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[15] + crit*ep))

```

```
razoes_3 = rbind(razoes_3,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))
```

```
# Versão 2:
```

```
(razao_est = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[16]))
(ep = sqrt(vcov_2.2[5,5] + vcov_2.2[16,16] + 2*vcov_2.2[5,16]))
(razao_li = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[16] - crit*ep))
(razao_ls = exp((-1)*betas_2.2[5] + (-1)*betas_2.2[16] + crit*ep))
razoes_3 = rbind(razoes_3,cbind(round(razao_est,2),paste('(',round(razao_li,2),' ' ;
+ ',round(razao_ls,2),')'))))
```

```
razoes_3
```


ANEXO B: Comandos para o ajuste do modelo misto - *software* SAS

```

proc sort data=dados_visao1 out=dados2;
by tipo descending bloco descending pos descending cod_estacao;
run;

/*Modelo 1: referência --> espaço vazio*/
proc glimmix data=dados2 asycov method=quad order=data;
title 'REFERÊNCIA: ESPAÇO VAZIO';
class tipo bloco cod_estacao pre pos ponto_experimento;
model tipo (ref='Espaço vazio') = bloco pre pos cod_estacao pre*cod_estacao
/ dist=multinomial link=glogit solution oddsratio cl covb;
random intercept / subject=ponto_experimento group=tipo solution;
output out=ajuste_misto pred=logito_ind pred(ilink)=prob_ind stderr(noblup)=ep_logito;
run;

proc sort data=ajuste_misto out=auxiliar(drop=tipo cod_tipo cod_tipo2 prob_ind ponto_piquete
estacao piquete combin);
by cod_estacao bloco pre pos ponto_experimento;
run;

proc transpose data=work.auxiliar out=auxiliar_2 prefix=logito;
by cod_estacao bloco pre pos ponto_experimento;
id _level_;
var logito_ind;
run;

proc transpose data=work.auxiliar out=auxiliar_3 prefix=ep;
by cod_estacao bloco pre pos ponto_experimento;
id _level_;
var ep_logito;
run;

data ic_mod1;
merge auxiliar_2 auxiliar_3;
by cod_estacao bloco pre pos ponto_experimento;
run;

data ic_mod1_2;
set ic_mod1 (drop='logitoEspaço vazio'n 'epEspaço vazio'n _label_ _name_);
rename logitoTouceira=logito1_1 'logitoPlanta invasora'n=logito2_1
epTouceira=ep1_1 'epPlanta invasora'n=ep2_1;
run;

data ic_mod1_2;
set ic_mod1_2;

```

```

li_logito1_1 = logito1_1 - 1.96*ep1_1;
ls_logito1_1 = logito1_1 + 1.96*ep1_1;

li_logito2_1 = logito2_1 - 1.96*ep2_1;
ls_logito2_1 = logito2_1 + 1.96*ep2_1;

prob_touc = exp(logito1_1)/(1 + exp(logito1_1) + exp(logito2_1));
prob_inv = exp(logito2_1)/(1 + exp(logito1_1) + exp(logito2_1));

li_prob_touc = exp(li_logito1_1)/(1 + exp(li_logito1_1) + exp(li_logito2_1));
ls_prob_touc = exp(ls_logito1_1)/(1 + exp(ls_logito1_1) + exp(ls_logito2_1));

li_prob_inv = exp(li_logito2_1)/(1 + exp(li_logito1_1) + exp(li_logito2_1));
ls_prob_inv = exp(ls_logito2_1)/(1 + exp(ls_logito1_1) + exp(ls_logito2_1));
run;

/*Modelo 2: referência --> touceira*/
proc glimmix data=dados2 asycov method=quad order=data;
title 'REFERÊNCIA: TOUCEIRA';
class tipo bloco cod_estacao pre pos ponto_experimento;
model tipo (ref='Touceira') = bloco pre pos cod_estacao pre*cod_estacao
/ dist=multinomial link=glogit solution oddsratio cl covb;
random intercept / subject=ponto_experimento group=tipo solution;
output out=ajuste_misto2 pred=logito_ind pred(ilink)=prob_ind stderr(noblup)=ep_logito;
run;

proc sort data=ajuste_misto2 out=auxiliar2(drop=tipo cod_tipo cod_tipo2 prob_ind ponto_piquete
estacao piquete combin);
by cod_estacao bloco pre pos ponto_experimento;
run;

proc transpose data=work.auxiliar2 out=auxiliar2_2 prefix=logito;
by cod_estacao bloco pre pos ponto_experimento;
id _level_;
var logito_ind;
run;

proc transpose data=work.auxiliar2 out=auxiliar2_3 prefix=ep;
by cod_estacao bloco pre pos ponto_experimento;
id _level_;
var ep_logito;
run;

data ic_mod2;
merge auxiliar2_2 auxiliar2_3;

```

```

by cod_estacao bloco pre pos ponto_experimento;
run;

data ic_mod2_2;
set ic_mod2 (drop='logitoTouceira'n 'epTouceira'n _label_ _name_);
rename 'logitoEspaço vazio'n=logito1_2 'logitoPlanta invasora'n=logito2_2
'epEspaço vazio'n=ep1_2 'epPlanta invasora'n=ep2_2;
run;

data ic_mod2_2;
set ic_mod2_2;

li_logito1_2 = logito1_2 - 1.96*ep1_2;
ls_logito1_2 = logito1_2 + 1.96*ep1_2;

li_logito2_2 = logito2_2 - 1.96*ep2_2;
ls_logito2_2 = logito2_2 + 1.96*ep2_2;

prob_touc = exp(logito1_2)/(1 + exp(logito1_2) + exp(logito2_2));
prob_inv = exp(logito2_2)/(1 + exp(logito1_2) + exp(logito2_2));

li_prob_touc = exp(li_logito1_2)/(1 + exp(li_logito1_2) + exp(li_logito2_2));
ls_prob_touc = exp(ls_logito1_2)/(1 + exp(ls_logito1_2) + exp(ls_logito2_2));

li_prob_inv = exp(li_logito2_2)/(1 + exp(li_logito1_2) + exp(li_logito2_2));
ls_prob_inv = exp(ls_logito2_2)/(1 + exp(ls_logito1_2) + exp(ls_logito2_2));
run;

/*Modelo de efeitos fixos (para testar o grau de dependência entre as observações)*/
proc glimmix data=dados2 asycov order=data;
title 'MODELO DE EFEITOS FIXOS';
class tipo bloco cod_estacao pre pos ponto_experimento;
model tipo (ref='Espaço vazio') = bloco pre pos cod_estacao pre*cod_estacao
/ dist=multinomial link=glogit solution oddsratio cl covb;
run;

```