JACKELYA ARAUJO DA SILVA

EQUAÇÕES DE ESTIMAÇÕES GENERALIZADAS PARA DADOS ORDINAIS EM ANÁLISE SENSORIAL DE CAFÉS ESPECIAIS E CRITÉRIOS DE SELEÇÃO PARA MATRIZES DE CORRELAÇÃO DE TRABALHO

LAVRAS – MG

2017

JACKELYA ARAUJO DA SILVA

EQUAÇÕES DE ESTIMAÇÕES GENERALIZADAS PARA DADOS ORDINAIS EM ANÁLISE SENSORIAL DE CAFÉS ESPECIAIS E CRITÉRIOS DE SELEÇÃO PARA MATRIZES DE CORRELAÇÃO DE TRABALHO

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Prof. Dr. Marcelo Ângelo Cirillo
Orientador

LAVRAS – MG
2017

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Silva, Jackelya Araujo da
Equações de estimações generalizadas para dados ordinais em análise sensorial de cafés especiais e critérios de seleção para matrizes de correlação de trabalho / . – Lavras : UFLA, 2017.
94 p. : il.

Tese(doutorado)–Universidade Federal de Lavras, 2017.
Orientador: Prof. Dr. Marcelo Ângelo Cirillo.
Bibliografia.

1. Análise Sensorial. 2. Dados correlacionados. 3. Cafés especiais. I. Silva, Jackelya Araujo da. II. Título.

JACKELYA ARAUJO DA SILVA

EQUAÇÕES DE ESTIMAÇÕES GENERALIZADAS PARA DADOS ORDINAIS EM ANÁLISE SENSORIAL DE CAFÉS ESPECIAIS E CRITÉRIOS DE SELEÇÃO PARA MATRIZES DE CORRELAÇÃO DE TRABALHO

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 11 de Julho de 2017.

Dra. Carla Regina Guimarães Brighenti	UFSJ
Dr. Flávio Meira Borém	DEG - UFLA
Dr. Júlio Silvio de Sousa Bueno Filho	DES - UFLA
Dra. Letícia Lima Milani Rodrigues	UNIFAL
Dr. Paulo Henrique Sales Guimarães	DES - UFLA

Prof. Dr. Marcelo Ângelo Cirillo
Orientador

LAVRAS – MG
2017

Aos meus pais Maria Araújo Linhares(Dona Remédios) e Cosme Damião, pelo amor e educação. Aos meus familiares e amigos. **DEDICO**

AGRADECIMENTOS

Ao Senhor da minha vida. Muito obrigada pela presença constante e marcante.

Aos meus pais, Maria Araujo Linhares da Silva e Cosme Damião da Silva, pelo amor, dedicação e comprometimento com a minha formação.

Aos meus familiares, irmãos e sobrinhos pela torcida e alegrias.

Aos meus amigos de longas datas. Agradeço pelas conversas noturnas e pela amizade.

Ao Professor Dr. Marcelo Ângelo Cirillo, que aceitou prontamente o convite para me orientar. Agradeço pela confiança a mim depositada, pela disponibilidade em discutir ideias e principalmente pela orientação e amizade. Por isso, muito obrigada.

Aos professores do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária pela contribuição da minha formação acadêmica. Agradeço à Nádia, pela prestatividade e amizade. Estranha!

Aos meus colegas e amigos conquistados em Lavras e vizinhanças. Os que são mineiros, muito obrigada por me ensinarem o "jeito" mineiro de ser. Em especial a uma mineira com quem convivi por dois anos, Carolina Bicalho. À ela, deixo o meu obrigada pela presença marcante da complexidade cultural existente no apartamento 104 da rua Waldemar Novaes. Aos nordestinos, agradeço por me fazerem lembrar da minha terra. Enfim, à todos que estiveram comigo nessa fase da minha vida. Muito obrigada.

Ao Lourenço Manuel pelo respeito, admiração, amizade e companherismo. Agradeço por me proporcionar momentos de alegria e descontração. Pela diversidade cultural e ao mesmo tempo tão igual.

À Universidade Federal do Piauí(UFPI) que autorizou o meu afastamento para que eu pudesse realizar o meu aperfeiçoamento profissional.

Ao Departamento de Bacharelado em Estatística da UFPI pela aprovação e contribuição para que eu realizasse as minhas atividades de forma tranquila e com êxito. Aos meus colegas e professores da Estatística. Obrigada.

Enfim, obrigada a todos que contribuíram, incentivaram e esperavam pelo término dessa jornada. Muito obrigada!

RESUMO

Neste trabalho estão presentes duas partes. A primeira parte contempla a fundamentação teórica desta tese. A segunda parte é composta de dois artigos científicos. O primeiro artigo, refere-se a modelagem em análise sensorial para múltiplas respostas repetidas em um experimento em análise sensorial, realizado com cafés especiais. A análise sensorial aplicada aos cafés especiais permitiu a organização de um conjunto de dados com medidas repetidas em níveis de provadores/genótipos ao longo de quatro safras. Isso ocorreu, devido ao fato de que diferentes provadores para diferentes amostras de cinco xícaras, realizaram avaliações de um mesmo genótipo em duas situações: ao longo das safras e durante a execução da degustação para atribuição das notas. Nesse sentido, houve a necessidade do estudo das associações em duas direções. A primeira no que se refere ao provador, e a segunda direção associada às notas ao efeito das safra. Concluiu-se que a metodologia proposta nesse primeiro artigo identificou as covariáveis sensoriais que são semelhantes ao longo das safras, bem como produziu estimativas de probabilidades para a categorização dos cafés especiais nas classes de melhores notas, associadas as degustações realizadas por safra. O segundo artigo apresenta um critério de seleção para matriz de correlação de trabalho, utilizada em equações de estimação generalizadas. O referido critério, diferentemente dos critérios de seleção expostos neste trabalho, faz uso da estimativa limitante dos parâmetros de associação como uma medida para a escolha da matriz de correlação de trabalho. Para tanto, realizou-se simulação Monte Carlo com diferentes cenários, comparando o seu resultado com os demais critérios. Além disso, são apresentadas duas aplicações, uma está relacionada a um conjunto de dados consagrados da literatura e a outra refere-se ao conjunto de dados provenientes de uma análise sensorial de cafés especiais. Foi possível concluir que o critério proposto, mostrou-se competitivo aos demais critérios.

Palavras-chave: Análise sensorial. Critério. Cafés especiais. Dados correlacionados. Matriz de correlação.

ABSTRACT

In this work two parts are presented. The first part considers the theoretical basis of this thesis. The second part is composed of two scientific articles. The first article refers to modeling in sensory analysis for multiple repeated responses in an experiment with specialty coffees. In the sensory analysis applied to specialty coffees, it was possible to construct a data set with repeated measurements at taster / genotype levels and over four crop seasons. This was due to the fact that different tasters for different cup tests carried out evaluations of the same genotype in two situations: throughout the crop seasons and during the execution of the tasting to assign the notes. In this sense, it was necessary to study the associations in two directions. The first one regarding the taster and the second direction associated with the grades to the effect of the harvest. It was concluded that the methodology proposed in this first article identified the sensory covariates that are similar throughout the harvests, as well as producing estimates of probability for the categorization of specialty coffees in the best grades classes, associated with tastings performed by harvest. The second article presents a selection criterion for labor correlation matrix, used in generalized estimation equations. This criterion, unlike the selection criteria presented in this paper, makes use of the limiting estimate of the association parameters as a measure for the choice of the work correlation matrix. For that, Monte Carlo simulation was performed with different scenarios, comparing its result with the other criteria. In addition, two applications are presented, one related to a set of literature data and the other refers to the set of data coming from a sensory analysis of specialty coffees. It was possible to conclude that the proposed criterion proved to be competitive to the other criteria.

Keywords: Sensory analysis. Criterion. Specialty coffee. Correlation data. Working correlation structure.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Na análise de dados com medidas repetidas, existe uma variedade considerável de técnicas quando a variável resposta segue uma distribuição normal: análise multivariada de perfis; análise de curvas de crescimento e modelos normais de efeitos aleatórios. Porém, não atentendo ao pressuposto de normalidade da variável resposta, uma série de dificuldades podem surgir devido à escassez de técnicas de análises que envolvam experimentos em análise sensoriais nas quais as respostas são pontuadas em uma escala de pontos entre zero e 10 pontos.

A qualidade sensorial é a última medida da qualidade de um produto. A análise sensorial compreende uma variedade de ferramentas poderosas e sensíveis para medir as respotas humanas à alimentos e outros produtos. A seleção do teste apropriado, condições de teste e análise de dados produzem resultados relevantes.

Basicamente, a aplicação da análise sensorial aos cafés especiais permite obter percepções específicas sobre os atributos sensoriais, bem como a identificação e interpretação dos componentes qualitativos que contribuem para o conceito final dos cafés especiais. Logo, as respostas produzidas em análise sensorial em uma escala entre zero e dez pontos, podem ser agrupadas em categorias, de modo a fornecer informação sobre o conjunto de atributos avaliados.

Considerando as respostas oriundas de um experimento em análise sensorial aplicada a diferentes genótipos de cafés, a avaliação para preferência e qualidade sensorial cafés especiais, é dada em termos de pontuações nas quais o conceito final para determinação da nota ao atributo, é formada por um conjunto de outras variáveis de aspectos qualitativos.

O conjunto de dados estudados neste trabalho é proveniente de um experimento em análise sensorial aplicada aos cafés especiais com medidas repetidas em níveis de provadores e genótipos repetidos ao longo de quatro safras. Uma das peculiaridades do experimento é que diferentes provadores realizaram avaliações para vários conjuntos de amostras de cinco xícaras de um mesmo genótipo em duas ocasiões: em quatro safras e durante a execução da degustação para atribuição das notas às amostras.

Dada a natureza da variável resposta categórica ordinal, há a necessidade do estudo das associações entre as categorias. Desse modo, as equações de estimação generalizadas(GEE) é uma abordagem que propõe analisar dados com medidas repetidas utilizando modelos lineares generalizados. Na metodologia GEE para dados ordinais a estimação dos parâmetros de associação representados por todos os pares possíveis de razão de chances, é uma medida obtida pelo

ajuste dos modelos marginais para os pares de respostas repetidas, que geralmente são baseados em probabibildades de respostas acumuladas, em vez de probabilidades das categorias.

Dessa forma, como a classificação dos conceitos finais aos atributos sensoriais possuem uma ordenação natural, para a metodologia GEE para dados ordinais, os modelos logits de probabilidades acumuladas incorporam esta ordenação indiretamente na sua construção.

Em se tratando da metodologia GEE, na sua formulação, faz-se uso de uma matriz simétrica, denominada de matriz de correlação de trabalho. Essa matriz pode possuir uma dentre as várias estruturas que constam na literatura por exemplo, pode assumir a estrutura permutável em que as associações para um mesmo indivíduo é considerada a mesma, ou uma estrutura auto-regressiva de ordem um (AR(1)), cuja característica da estrutura de correlação é que a magnitude das correlações (positivas) diminui rapidamente ao longo do tempo e a separação entre os pares de medidas repetidas aumenta. Mas, a tarefa de identificar quais dentre as várias estruturas utilizar para iniciar o processo iterativo para obtenção das estimativas dos parâmetros de regressão, de modo a manter as propriedades de consistência e eficiência das estimativas, fica sob escolha do pesquisador que deve levar em consideração as características do estudo.

No entanto, critérios estatísticos que auxiliem na escolha da matriz de correlação de trabalho, têm sido objetos de estudos, pois entende-se que a seleção da matriz de correlação de trabalho quando melhor especificada na metodologia GEE, evitará a perda da eficiência das estimativas dos parâmetros do modelo marginal, bem como preservará as condições de consistência dos parâmetros de associação.

Mediante ao exposto, este trabalho tem por objetivo realizar um estudo para um conjunto de dados provenientes de uma análise sensorial aplicada aos cafés especiais, considerando as associações entre as degustações ao longo das safras e propor um novo critério de seleção para a escolha da matriz de correlação de trabalho.

Dentre os objetivos específicos destacam-se:

a) Fornecer uma estratégia de modelagem para os estudos das avaliações das associações em duas direções;

b) Comparar o desempenho do critério de seleção para a matriz de correlação de trabalho com relação aos demais critérios da literatura, em diferentes cenários via simulação Monte Carlo;

O trabalho está disposto em formato de artigo, sendo constituído por duas partes:

- A primeira parte é composta de uma introdução geral, dos objetivos e em seguida é exposto o referêncial teórico, base para a fundamentação do que é apresentado nos artigos, que compõem a segunda parte deste trabalho.
- A segunda parte é constituída por dois artigos:
    i. O artigo 1 que consiste em apresentar uma estratégia de modelagem para as associações entre as respostas dentro dos grupos formados por provadores e genótipos, e para a concordância das notas fornecidas pelos provadores aos cafés especiais avaliadas por safra, em conjunto com os atributos qualitativos. Os principais aspectos dos resultados foram discutidos.
    ii. O artigo 2 que consiste em apresentar um critério de seleção para matriz de correlação de trabalho. O critério proposto, JCC, foi comparado em relação a alguns critérios existentes da literatura, e o desempenho em termos de proporções foi discutido para dados normais e binários.
- As considerações finais são apresentadas ao final da segunda parte que compõe o corpo desta tese, bem como os aspectos relevantes deste trabalho para os estudos em análise sensorial e a contribuição em estudos relacionados às equações de estimação generalizadas.

2 REFERENCIAL TEÓRICO

Serão apresentadas inicialmente nessa seção as principais características sobre os atributos sensoriais avaliados, bem como a abordagem de equações de estimação generalizadas (GEE). Em seguida, serão discutidas a importância da matriz de correlação de trabalho para a metodologia GEE, e posteriormente apresentou-se alguns dos critérios para a seleção da matriz de correlação de trabalho e suas formulações.

2.1 Análise sensorial e atributos sensoriais avaliados para classificação de cafés especiais

Segundo Schmidt e Miglioranza (2011), o café foi cultivado pela primeira vez pelos árabes, por isso a denominação *Coffea arabica L.*, nome científico da mais importante espécie. O café Arábica (*Coffea arabica L.*) representa cerca de dois terços da produção mundial (ILLY, 2002).

A qualidade da bebida do café pode ser medida pela satisfação dos consumidores e está associada ao sabor e aroma com o qual este se apresenta. De acordo com Malavolta (2000), a qualidade do café refere-se ao conjunto de características sensoriais do grão ou da bebida que imprimem a este um valor comercial.

Os estudos da análise sensorial do café têm evoluido e tornou-se indispensável para a indústria de alimentos e dispõe de vários métodos distintos, (discriminativos, descritivos e afetivos), utilizados por diferentes tipos de provadores para avaliação das amostras desgustadas (SCHMIDT; MIGLIORANZA, 2011).

Após aperfeiçoamentos, a Associação Americana de Cafés especiais - SCAA elaborou um protocolo capaz de avaliar dez diferentes atributos (Fragrância/Aroma, Uniformidade, Defeitos, Doçura, Sabor, Acidez, Corpo, Xícara limpa, Harmonia e Impressão global), sendo que cada um é pontuado numa escala entre zero e dez. O café especial é aquele que atinge nota final acima de 80 pontos(SCAA, 2015).

Ainda, segundo o protocolo da SCAA (SCAA, 2015) para o procedimento das análises sensoriais dos cafés especiais, é necessário pelo menos a realização de cinco xícaras de cada amostra no teste sensorial para as avaliações dos aspectos específicos observados como a qualidade e intensidade. Os registros das avaliações tem por objetivo determinar as diferenças sensoriais reais entre as amostras, descrever o sabor e determinar a preferência. A seguir, na Tabela 2.1, uma breve descrição dos atributos avaliados nas provas de xícaras.

Tabela 2.1 – Atributos sensoriais avaliados nas provas de xícaras

Atributos sensoriais	Descrição
Fragância/Aroma	Os aspectos aromáticos incluem fragância (Definido como cheiro do café moído, ainda seco) e Aroma (Definido como cheiro do café quando diluído em água quente)
Sabor	Personagem principal. É uma impressão combinada de todas as sensações gustativas. Abrange a complexidade da combinação dos gostos básicos (doce, salgado, amargo e ácido)
Impressão Final	O sabor final é definido como o comprimento do sabor positivo que permanece depois que o café foi degustado.
Acidez	Contribui para o caráter de doçura e frescura do café. Está relacionado ao tipo de acidez, se é desejável ou não.
Corpo	A qualidade do corpo é baseada na sensação tátil do líquido na boca, percebida entre a língua e o palato. Está relacionado a textura e densidade do café.
Harmonia	Combinação de todos os atributos.
Doçura	Refere-se ao sabor agradável. Gosto básico, muito apreciado na bebida.
Uniformidade	Consistência dos mesmos atributos em todas as xícaras das mesmas amostra avaliadas.
Defeitos	Refere-se aos sabores negativos na bebida. Ocasiona diminuição na pontuação final.
Xícara limpa	Comprovação de que o café está livre de defeitos.

Fonte: Adaptado de SCAA (2015)

A prova de xícara depende do treinamento, ou mesmo da frequência com que os provadores realizam as degustações de determinados tipos de cafés. Eles podem desenvolver habilidades sensoriais distintas o que acarreta distorções, fazendo com que haja discordância entre as notas dadas para as amostras provadas por diferentes provadores(MAZZAFERA et al., 2002).

Diversos fatores podem influenciar a composição química do grão, e consequentemente a qualidade do café produzido. Destacam-se as características genéticas, ambientais e culturais (CHAGAS; MALTA; PEREIRA, 2005).

Segundo Borém et al. (2008) a qualidade final do café é definida por um conjunto de atributos que irão depender da espécie, variedade, solo e ambiente de produção; época e método de colheita, processamento e secagem entre outros, até a disponibilização ao consumidor.

Contudo, os cafés de boa qualidade exigem tratos especiais desde a fase de pré-colheita, passando pela colheita, até a pós-colheita, eliminando, assim, possíveis fatores que possam interferir da qualidade da bebida futuramente (BORÉM, 2008).

2.2 Equações de estimação generalizadas (GEE)

Serão apresentadas a metodologia GEE para dados contínuos e binários. Segue também a introdução da notação utilizada neste trabalho, bem como apresentação dos métodos GEE1 e GEE2 para dados ordinais. A metodologia GEE2 e ALR para dados ordinais não serão utilizadas neste trabalho, porém a apresentação será feita, pois são metodologias de equações de estimação generalizadas para dados ordinais presentes na literatura. Para completar essa seção, será apresentada a metodologia GEE utilizando a medida kappa.

2.2.1 Equações de estimação para variáveis contínuas e binárias

Introduzido por Nelder e Wedderburn (1972), os modelos lineares generalizados é uma abordagem que corresponde a uma síntese de modelos desenvolvidos para fazer face a situações de natureza experimental ou observacional, que não eram adequadamente explicadas pelo modelo linear normal. Alguns deles são, os modelos *probit*, *complemento log-log* e *logit*.

Modelos Lineares Generalizados baseiam-se na família exponencial de distribuição de probabilidade, que inclui a distribuição normal, binomial, poisson, gama, gaussiana inversa e geométrica. Com base em verossimilhança os modelos lineares generalizados assumem que os indivíduos sejam independentes. No entanto, no caso de dados agrupados, essa suposição pode não ser atendida. Assim, Liang e Zeger (1986) introduziram a metodologia GEE, a qual foi explicitamente desenvolvida para servir como método para ampliar os modelos lineares generalizados para dados correlacionados.

Em síntese, a abordagem GEE é aplicada a análise de dados com medidas repetidas utilizando modelos lineares generalizados em que se assume i sujeitos independentes em um experimento que são observados em t ocasiões. A metodologia GEE não pressupõe a especificação completa da distribuição multivariada das respostas repetidas, porém requer a identificação dos dois primeiros momentos(LIANG; ZEGER, 1986).

A análise sob abordagem GEE pode ser escolhida a partir de três diferentes métodos para se estimar os parâmetros de regressão β e os parâmetros de associação α. O primeiro método é conhecido como equações de estimação generalizadas de primeira ordem (GEE1), que trata os parâmetros α como parâmetro de perturbação e cujo interesse principal está na obtenção das estimativas de β (LIANG; ZEGER, 1986).

O segundo método, proposto por Prentice e Zhao (1991) é denominado de GEE2, e utiliza equações de estimação para obtenção das estimativas dos parâmetros de regressão e de associação conjuntamente. Essa abordagem permite estimar os parâmetros de associação α mais precisamente, porém existe a desvantagem de que a consistência dos parâmetros de regressão β depende da especificação correta do modelo além de grande esforço computacional.

O terceiro método, denominado de regressão logística alternada (ALR) é uma abordagem alternativa para a modelagem da média marginal e para a estimação dos parâmetros de associação, envolvendo covariáveis, e faz uso de pares de razão de chances. O algoritmo de estimação dos parâmetros alterna entre a regressão logística usando GEE1, e uma outra regressão logística para cada uma das outras respostas de associação de um mesmo grupo (CAREY; ZEGER; DIGGLE, 1993).

Dada a composição desta tese, inicia-se a introdução da notação a ser utilizada na primeira parte deste trabalho e posteriormente apresenta-se as notações para a segunda parte que compõe o corpo desta tese. Para tanto, denotaremos por $\mathbf{Y}_i = \{Y_{i1}, \ldots, Y_{in_i}\}$ o vetor de respostas normais ou binárias para o $i-$ésimo indivíduo, $i = 1, 2, \ldots, K$ observado nos tempos $t = 1, 2, \ldots, n_i$. E seja, $\mathbf{X}_i = \{x_{i1}, \ldots, x_{in_i}\}$ a matriz de covariáveis, $n_i \times p$, com p variáveis explicativas associadas ao $i-$indivíduo. A princípio, assume-se que é conhecida a distribuição marginal de Y_{it}, dada por

$$f(y_{it}) = exp[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}\phi], \tag{2.1}$$

em que $\theta_{it} = h(\eta_{it})$, $h^{-1}(\theta_{it})$ é a função de ligação, a e b são funções reais, respectivamente de θ_{it} e y_{it}, $\eta_{it} = x_{it}^T\beta$ é o preditor linear. E por essa formulação, o primeiro e segundo momentos de Y_{it} são dados por

$$E(Y_{it}) = a'(\theta_{it}), \qquad var(Y_{it}) = a''(\theta_{it})/\phi.$$

Sejam $\mathbf{C}_i(\rho)$ e $\Sigma_i(\rho)$ as respectivas matrizes $n_i \times n_i$ de correlações e covariâncias verdadeiras de $\mathbf{Y}_i$, que usualmente são desconhecidas. Aqui, ρ é um vetor de parâmetros de correlação que caracterizam completamente $\mathbf{C}_i(\rho)$. Para o caso em que $\mathbf{C}_i(\rho)$ é desconhecida, Liang e Zeger (1986), introduziram as equações de estimação generalizadas para a obtenção das estimativas dos parâmetros β, baseadas em uma matriz simétrica, $n_i \times n_i$, denominada de matriz de correlação de trabalho $\mathbf{R}_i(\alpha)$, α é um vetor de parâmetros de correlação, que na prática, também é desconhecida (SUTRADHAR; DAS, 2000).

Assim, denotaremos por $\hat{\beta}_G$, as estimativas para os parâmetros β, e $\hat{\mathbf{V}}_G$ a matriz de covariância estimada, provenientes do processo iterativo para obtenção das estimativas dos parâmetros β, sob a suposição de uma matriz de trabalho quaisquer.

Neste trabalho, serão tratados somente três estruturas para a matriz de correlação de trabalho. A estrutura permutável em que se pressupõe que a correlação entre quaisquer pares de medidas de um mesmo indivíduo é a mesma, $\alpha, \forall t \neq t'$. É frequentemente usada como escolha prática em pequenas amostras. A estrutura independente que é uma matriz identidade, então não há parâmetro de associação para os pares de respostas. E a matriz de correlação de trabalho auto-regressiva de ordem um (AR(1)), cuja característica da estrutura de correlação é que a magnitude das correlações (positivas) diminui rapidamente ao longo do tempo e a separação entre os pares de medidas repetidas aumenta($\alpha_{tt'} = \alpha^{|t-t'|}$). Em analogia a estrutura de correlação auto-regressivade ordem um, Verbeke (2005) sugere o modelo $log(\theta) = \frac{1}{|t'-t|}\alpha, t' \neq t$. Os valores de θ indicam a direção da associação entre as respostas, quando $\theta = 1$ indica independência, $\theta < 1$, associação negativa e para $\theta > 1$ correponde a associação positiva.

No que segue, Liang e Zeger (1986) assumiram que $\hat{\alpha}$ é um estimador consistente de α, e sugeriram método dos momentos para sua obtenção, tais que $\mu_{it} = g(x_{it}^T\beta)$ e variância $\phi\sigma_{it}^2$, sendo ϕ um parâmetro de escala desconhecido que geralmente assume-se $\phi = 1$, $\mu_i = (\mu_{it})$ será o vetor de médias marginais, $\mathbf{A}_i = diag(\sigma_{it}^2)$ e a matriz de covariância pode ser escrita como $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}$. As estimativas dos parâmetros β, será a solução de:

$$\mathbf{U}(\beta, \alpha) = \sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mu_i) = \mathbf{0}, \tag{2.2}$$

em que $\mathbf{D}_i = \partial\mu_i/\partial\beta$.

Para obtenção das estimativas $\hat{\beta}$, realiza-se iteração entre escore de Fisher modificado para estimação dos parâmetros β, e método dos momentos para obtenção das estimativas $\hat{\alpha}$. E

assim, dado $\hat{\alpha}$, as estimativas para β podem ser obtidas pelo processo iterativo:

$$\hat{\beta}_{m+1} = \hat{\beta}_m - \left(\sum_{i=1}^{K} \mathbf{D}_i^T(\hat{\beta}_m) \tilde{\mathbf{V}}_i^{-1}(\hat{\beta}_m) \mathbf{D}_i(\hat{\beta}_m) \right)^{-1} \left(\sum_{i=1}^{K} \mathbf{U}_i(\hat{\beta}_m, \hat{\alpha}\{\beta, \hat{\phi}(\beta)\}) \right), \tag{2.3}$$

em que $\tilde{\mathbf{V}}_i(\hat{\beta}_m) = V_i[\beta, \hat{\alpha}, \hat{\phi}(\beta)]$. Consequentemente, a estimação dos parâmetros α pode ser realizada por meio dos resíduos de Pearson definidos por:

$$\hat{\varepsilon}_{it} = (Y_{it} - a'(\theta_{it}))/\sqrt{a''(\theta_{it})} \tag{2.4}$$

tal que, um estimador natural para $\alpha = (\alpha_1, \ldots, \alpha_{n-1})^T$, dado as estimativas $\hat{\beta}$, com $\alpha_t = corr(Y_{it}, Y_{i,t+1})$ para $t = 1, \ldots, n_i - 1$

$$\hat{\alpha}_t = \phi \frac{1}{K-p} \sum_{i=1}^{K} \hat{\varepsilon}_{it} \hat{\varepsilon}_{i,t+1} \tag{2.5}$$

de modo que para as estruturas de correlações permutável, $\alpha = corr(Y_{it}, Y_{i,t'})$, $t \neq t'$ e AR(1), os parâmetros de correlação, α, podem ser estimados, respectivamente por (WANG; CAREY, 2003):

$$\hat{\alpha}(\hat{\beta}) = \frac{\sum_{i=1}^{K} \sum_{t>t'} \hat{\varepsilon}_{it} \hat{\varepsilon}_{it'}}{\sum_{i=1}^{K} (n_i - 1) \sum_{t=1}^{n_i} \hat{\varepsilon}_{i,t}} \tag{2.6}$$

$$\hat{\alpha}(\hat{\beta}) = \frac{\sum_{i=1}^{K} \sum_{t=2}^{n_i} \hat{\varepsilon}_{it} \hat{\varepsilon}_{i,t-1}}{\sum_{i=1}^{K} \left\{ \sum_{t=2}^{n_i - 1} \hat{\varepsilon}_{it}^2 + (1/2)(\hat{\varepsilon}_{i1}^2 + \hat{\varepsilon}_{in_i}^2) \right\}} \tag{2.7}$$

2.2.2 Equações de estimação generalizadas para dados ordinais

O modelo GEE no qual se baseia uma das propostas deste trabalho, tem como fundamentação a abordagem realizada por Heagerty e Zeger (1996), e pela proposta de extensão para dados longitudianais de Williamson, Kim e Lipsitz (1995), em que consideram razão de chances para medir a associação entre cada observação para um mesmo grupo.

Williamson, Kim e Lipsitz (1995) em um estudo oftalmológico, introduziram uma classe de equações de estimação generalizadas para análise de dados bivariados. Consideraram o fato de que a resposta de interesse em seu estudo podiam ser obtidas para cada um dos olhos dos indivíduos, como também ser descrita pelo indivíduo. Produzindo assim, respostas repetidas entre olhos de um mesmo indivíduo.

A relação entre as respostas bivariadas foram descritas usando pares de razão de chances para todas as combinações possíveis, razão de chances global, de categorias ordinais assumidas no estudo.

Diferentemente do proposto por Liang e Zeger (1986), Prentice e Zhao (1991) e Carey, Zeger e Diggle (1993) para o processo de estimação dos parâmetros de associação, α, Williamson, Kim e Lipsitz (1995) introduziram um segundo conjunto de equações de estimação para obtenção das estimativas $\hat{\alpha}$. Para tanto, em seu estudo ofalmológico, considerou K indivíduos com T_i tempos de observações para cada um dos olhos avaliados. E associados a cada um dos olhos dos indivíduos no $t-$ésimo tempo, sejam as covariáveis $\mathbf{X}_{it}$, e denotaram por Z_{it} as respostas categóricas, $k = 1,2,\ldots,c-1$, para os olhos do $i-$ésimo indivíduo observado no $t-$ésimo tempo. Definiram, portanto a variável aleatória

$$Y_{itk} = \begin{cases} 1, & \text{se} \quad Z_{it} = k \\ 0, & \text{caso contrário.} \end{cases} \tag{2.8}$$

As respostas categóricas a que se refere é associada à identificação dos fatores de risco para a retinopatia diabética. A gravidade da retinopatia diabética foi classificada de acordo com uma escala ordinal de 10 pontos. Combinaram a escala original para formar categorias ordenadas de nenhuma, leve, moderada e proliferativa.

Para as repostas marginais, $\gamma_{itk} = P(Z_{it} \leq k|\mathbf{X}_{it} = x_{it})$ consideraram a função de ligação $g(.)$ e definiram o modelo como

$$g(\gamma_{itk}) = \theta_k + \mathbf{x}_{it}^T\beta. \tag{2.9}$$

Desse modo, para a função de distribuição bivariada, $F_{ijk} = P(Z_{i1} \leq j, Z_{i2} \leq k)$, $j,k = 1,\ldots,c$ para $i = 1,2,\ldots,K$, a razão de chances global para o $i-$ésimo indivíduo em termos de F_{ijk}, γ_{i1j} e γ_{i2k} é dada por:

$$\psi_{ijk} = \frac{F_{ijk}(1-\gamma_{i1j}-\gamma_{i2k}+F_{ijk})}{(\gamma_{i1j}-F_{ijk})(\gamma_{i2j}-F_{ijk})}. \tag{2.10}$$

Dada as especificações, sejam $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{iT_i})$, em que $\mathbf{Y}_{it} = (Y_{it1}, \ldots, Y_{it,c-1})$ e $\pi_i = (\pi_{i1}, \ldots, \pi_{iT_i})$ tal que $E(Y_{itk}) = \pi_{itk}(\beta) = \gamma_{itk} - \gamma_{it,k-1}$.

O primeiro conjunto de equações de estimação para as médias marginais é

$$v_1(\beta, \alpha) = \sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \pi_i(\beta)\} = \mathbf{0}, \tag{2.11}$$

em que $\mathbf{D}_i = \partial \pi_i(\beta)/\partial \beta$ e para ambos os olhos avaliados, a matriz de covariância de trabalho, $\mathbf{V}_i$ é uma matriz bloco,

$$\mathbf{V}_i = \begin{bmatrix} \mathbf{V}_{11i} & \mathbf{V}_{12i} \\ \mathbf{V}_{21i} & \mathbf{V}_{22i} \end{bmatrix}, \tag{2.12}$$

em que para o olho esquerdo, $\mathbf{V}_{11i} = Diag(\pi_{1i}) - \pi_{1i}\pi'_{1i}$ é uma matriz de covariância de dimensões $(c-1) \times (c-1)$, e de forma similar para o olho direito, $\mathbf{V}_{22i}$. As matrizes fora da diagonal principal representam as covariância entre os dois olhos, de modo que os elementos de $\mathbf{V}_{12i}$ e $\mathbf{V}_{11i}$ são $cov(Y_{i1j}, Y_{i2k}) = P(Y_{i1j} = 1, Y_{i2k} = 1) - P(Y_{i1j} = 1)P(Y_{i2k} = 1) = \omega_{ijk} - \pi_{i1j}\pi_{i2k}$.

O segundo conjunto de equações de estimação é desenvolvido da seguinte forma: para cada indivíduo considerou-se as variáveis indicadoras $U_{ijk} = I\{Y_{i1j} = 1, Y_{i2k} = 1\}$, $E(U_{ijk}) = E(Y_{i1j}Y_{i2k}) = \omega_{ijk}$ de modo que $\mathbf{U}_i$ é um vetor de $c^2 - 1 \times 1$,

$$\mathbf{U}_i = (U_{i11}, U_{i12}, \ldots, U_{i1c}, \ldots, U_{i21}, \ldots, U_{ic,c-1})$$

e de forma similar para $E(\mathbf{U}_i) = \omega_i(\beta, \alpha) = (\omega_{i11}, \ldots, \omega_{i1c}, \ldots, \omega_{i21}, \ldots, \omega_{ic,c-1})$.

Portanto, o segundo conjunto de equações de estimação é definido

$$v_1(\beta, \alpha) = \sum_{i=1}^{K} \mathbf{C}_i^T \mathbf{W}_i^{-1} \{\mathbf{U}_i - \omega_i(\beta, \alpha)\} = \mathbf{0}, \tag{2.13}$$

em que $\mathbf{W}_i$ é a matriz de covariância de trabalho de $\mathbf{U}_i$, e como U_{ijk} são variáveis binárias, então a matriz diagonal, $\mathbf{W}_i$, será composta dos elementos $\omega_{ijk}(1 - \omega_{ijk})$ e $\mathbf{C}_i = \partial \omega_i(\beta, \alpha)/\partial \alpha$.

Para computar $(\hat{\beta}, \hat{\alpha})$, o procedimento é semelhante às estimativas $\hat{\beta}$ para as equações (2.2), tais que

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} - \left(\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i\right)^{-1} \left(\sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \pi_i(\hat{\beta}^{(m)})\}\right) \tag{2.14}$$

e

$$\hat{\alpha}^{(m+1)} = \hat{\alpha}^{(m)} - \left(\sum_{i=1}^{K} \mathbf{C}_i^T \mathbf{W}_i^{-1} \mathbf{C}_i\right)^{-1} \left(\sum_{i=1}^{K} \mathbf{C}_i^T \mathbf{W}_i^{-1} \{\mathbf{U}_i - \omega_i(\hat{\beta}^{(m+1)}, \hat{\alpha}^{(m)})\}\right) \tag{2.15}$$

Note que na formulação do segundo conjunto de equações de estimação, na definição das variáveis indicadoras, U_{ijk}, é necessário que os indivíduos inseridos no estudo tenham ambos os olhos, avaliados no $t-$ésimo tempo de observação. Com isso, as equações de estimação descrita por Williamson, Kim e Lipsitz (1995), não são aplicadas para análise de dados cujos números de observações sejam variados. Assim, a abordagem GEE para o caso bivariado foi extendido para os grupos correlacionados, nos quais podem conter diferentes números de observações.

Nesse caso, o desenvolvimento da metodologia GEE para variados números de observações, ocorre considerando o grupo de indivíduos, nas quais as respostas repetidas são respostas dos indivíduos em diferentes ocasições. Dessa forma, associando ao estudo oftalmológico, as respostas bivaridas de um mesmo indivíduo serão avaliadas ao longo do tempo. Com isso, o interesse do estudo das associações estará relacionado às respostas do grupo ao longo do tempo e não mais, entre as respostas bivariadas de um mesmo indivíduo.

Williamson, Kim e Lipsitz (1995) considerando a abordagem em estudos longitudinais, reescreveu o caso das análise das associações entre as respostas do olho direito e esquerdo, tal que sua breve formulação é dada a seguir.

Seja $\psi_{ijk}(s,t)$ a razão de chances global do $i-$ésimo indivíduo com resposta na categoria j na $s-$ésima ocasião, e a resposta na categoria k na $t-$ésima ocasião de observação. Denota-se $F_{ijk}(s,t) = P(Z_{is} \leq j, Z_{it} \leq k)$ a distribuição acumulada conjunta para duas as ocasiões de observações. Assim, a equação (6) para as respostas longitudinais é dada por:

$$\psi_{ijk}(s,t) = \frac{F_{ijk(s,t)}\{1 - \gamma_{isj} - \gamma_{itk} + F_{ijk}(s,t)\}}{\{\gamma_{isj} - F_{ijk}(s,t)\}\{\gamma_{itk} - F_{ijk}(s,t)\}}, \tag{2.16}$$

para $i = 1,2,\ldots,K$, $s,t = 1,2,\ldots,T_i$ $(s \neq t)$ e $j,k = 1,2,\ldots,c-1$. Contudo, $\mathbf{Y}_i$ é um vetor de dimensão $T_i(c-1) \times 1$, $\mathbf{U}_i$ terá dimensão $T_i(T_i-1)(c^2-1)/2 \times 1$ e $\mathbf{W}_i$ será uma matriz de blocos

diagonais, $T_i(T_i-1)(c^2-1)/2 \times T_i(T_i-1)(c^2-1)/2$. E conforme descrito na equação (2.13), $\mathbf{W}_i$ é a matriz de covariância de trabalho de $\mathbf{U}_i$ e especificamente terá a seguinte estrutura:

$$\mathbf{W}_i = \begin{bmatrix} \mathbf{W}_{12i} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{13i} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{W}_{T_i-1,T_i,i} \end{bmatrix}, \tag{2.17}$$

para $i = 1,2,\ldots,K$, $s = 1,2,\ldots,T_i-1$ e $t = 2,3,\ldots,T_i (s \neq t)$.

O procedimento de estimação para os parâmetros de regressão, β, e de associação α ocorre da mesma forma que nas equações (2.14) e (2.15).

Seguindo a referência para modelagem GEE para dados ordinais, Heagerty e Zeger (1996), também consideraram a razão de chances para medir a associação entre cada observação para um mesmo grupo. Porém, diferentemente do apresentado por Williamson, Kim e Lipsitz (1995), os autores sugerem modelar a correlação entre pares de categorias distintas através de um modelo linear generalizado em função da estimativas dos parâmetros de associação α. Para tanto, seja O_i o vetor de medidas ordinais para o $i-$ésimo indivíduo, e que O_{it} representa a $t-$ésima observação do $i-$ésimo indivíduo, e x_{it} as covariáveis associadas as respostas ordinais O_{it}.

A medida ordinal $O_{it} = k$, em que $k \in \{1,2,\ldots c\}$ com c categorias de respostas, corresponde a um vetor de variáveis indicadoras acumuladas

$$Y_{it(k)} = I_{(O_{it}>k)}, \tag{2.18}$$

em que $k = 1,2,\ldots,c-1$.

A variável $Y_{it(k)}$ correspondente ao indivíduo i, avaliado do tempo t na $k-$ésima categoria, é uma variável binária e o modelo de razão de chances proporcionais para as médias marginais é dado por:

$$logit[E(Y_{it(k)})] = \theta_k + x_{it}^T \beta. \tag{2.19}$$

Para cada resposta O_{it}, associa-se um vetor x de p covariáveis x_{pt}, de modo que, fixado o $i-$ésimo indivíduo, $x = (x_1,x_2,\ldots,x_p)^T$ indica o vetor de covariáveis observadas em cada tempo ou ocasião de observação, $t = 1,2,\ldots,n_i$. Assim, o vetor de respostas para o indivíduo

i, na $t-$ésima observação, Y_{it}^t, segue uma distribuição Bernoulli com média $\mu_{it} = P(Y_{it} = 1)$. Logo, o vetor de respostas binárias para o $i-$indivíduo é dado por $Y_i = \{Y_{i1}^t, Y_{i2}^t, \ldots, Y_{it}^t\}^T$, e $\mu_i = E(Y_i)$.

Note que, para as categorias de respostas k, $(k = 1,2,\ldots,c)$ associadas às covariáveis, $x_i = (x_{1t}, x_{2t}, \ldots, x_{pt})$, o vetor de respostas para a $k-$ésima categoria, $Y_{it(k)}^t$, terá distribuição binomial com probabilidade de sucesso $\pi_k(x_i)$.

Para compreensão da estruturas das respostas $Y_{it(k)}$, segue a Tabela 2.2 como exemplo, em que se consideram três categorias de respostas, $(k = k_1, k_2, k_3)$ e fictícios valores observados para as respostas ordinais O_{it}.

Tabela 2.2 – Exemplo de respostas Y_i para 3 categorias de respostas para n indivíduos com n_i observações.

Indivíduos (i)	Observação (t)	Resposta ordinais O_{it}	Indicadoras $(k = 1,2)$	Valores de Y_{it}	Respostas Y_i
1	1	3	$\left(I_{(3>1)}, I_{(3>2)}\right)$	$(1,1)$	$(1,1)^t$
1	2	2	$\left(I_{(2>1)}, I_{(2>2)}\right)$	$(1,0)$	$(1,0)^t$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	n_1	2	$\left(I_{(2>1)}, I_{(2>2)}\right)$	$(1,0)$	$(1,0)^t$
2	1	1	$\left(I_{(1>1)}, I_{(1>2)}\right)$	$(0,0)$	$(0,0)^t$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	1	3	$\left(I_{(3>1)}, I_{(3>2)}\right)$	$(1,1)$	$(1,1)^t$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	n_K	1	$\left(I_{(1>1)}, I_{(1>2)}\right)$	$(0,0)$	$(0,0)^t$

A razão de chances para o par O_{it} e O_{il}, é um modelo de razão de chances proporcional que pode ser visualizado como uma regressão logística conjunta para cada uma das possíveis respostas binárias $Y_{it(k)}$, definida como:

$$\psi_{i(t,l)(k_1,k_2)} = \frac{P(O_{it} > k_1, O_{il} > k_2)P(O_{it} \leq k_1, O_{il} \leq k_2)}{P(O_{it} > k_1, O_{il} \leq k_2)P(O_{it} \leq k_1, O_{il} > k_2)}, \tag{2.20}$$

em que k_1 e k_2 categorias quaisquer.

Conforme definido em (2.18) e utilizando $\psi_{i(t,l)}$ como medida de associação das respostas dentro de um mesmo grupo, a expressão para a razão de chances global entre as resposta binárias $Y_{it(k_1)}$ e $Y_{it(k_2)}$, nomeada por $OR(Y_{it(k_1)}, Y_{it(k_2)})$, é estimada por:

$$logOR(Y_{it(k_1)}, Y_{it(k_2)}) = log\left(\frac{P(Y_{it} = 1, Y_{il} = 1)P(Y_{it} = 0, Y_{il} = 0)}{P(Y_{it} = 1, Y_{il} = 0)P(Y_{it} = 0, Y_{it} = 1)}\right). \tag{2.21}$$

Para a especificação do modelo marginal proposto por Heagerty e Zeger (1996), a correlação entre as respostas para os modelos de regressão de razão de chances definida como

$$\rho_{i(t,l)(k_1,k_2)}(\alpha) = Corr(Y_{it(k_1)}, Y_{il(k_2)}|X_{itl}) = \frac{exp(X_{itl}\alpha)-1}{exp(X_{itl}\alpha)+1},$$

é a correlação para as variáveis binárias, conforme definidas em (2.18). A correlação é obtida em função do vetor de parâmetros α, na qual a estrutura de correlação para as múltiplas respostas, depende de covariávies X_{itl} através da função de ligação $g(\rho_{i(t,l)}) = X_{itl}\alpha$ pelo seguinte modelo linear generalizado

$$log\left(\frac{1+\rho_{i(t,l)(k_1,k_2)}}{1-\rho_{i(t,l)(k_1,k_2)}}\right) = z^t_{i(t,l)(k_1,k_2)}\alpha, \quad i = 1,\dots K, \quad t,l = 1,\dots,n_i. \tag{2.22}$$

em que z é um subconjunto de (x_{it}, x_{il}) ou qualquer outra covariável relevante para modelar o grau de associação entre as t e $l-$ésima observações.

Dessa forma, as expressões (2.21) e (2.22) são dadas para quantificar a associação entre as observações t e l em relação ao $i-$ésimo indivíduo, para cada uma das categorias como:

$$logOR(Y_{it(k_1)}, Y_{il(k_2)}) = log\left(\frac{1+\rho_{i(t,l)}}{1-\rho_{i(t,l)}}\right) = X_{itl}\alpha, \quad i = 1,\dots K, \quad t,l = 1,\dots,n_i. \tag{2.23}$$

Heagerty e Zeger (1996) para utilização do método ALR (CAREY; ZEGER; DIGGLE, 1993), propuseram a utilização de pares de razão de chances como medida de associação para respostas binárias, conforme descrito:

$$logit[E(Y_{it(k_1)}|Y_{il(k_2)})] = \gamma_{itl}Y_{il(k_2)} + log(\Delta_{itl}) \tag{2.24}$$

e

$$\Delta_{itl} = \frac{\mu_{it} - \nu_{itl}}{1-\mu_{it}-\mu_{il}+\nu_{itl}},$$

em que, γ_{itl} é o log da razão de chances entre $Y_{it(k_1)}$ e $Y_{it(k_2)}$ e $\nu_{itl} = E(Y_{it(k_1)}Y_{il(k_2)})$ e definiram um conjunto de equações de estimação baseado em resíduos condicionais para calcular as estimativas dos parâmetros de associação. A princípio, construíram pares de produtos, Y_i^* e Y_i^{**},

tais que

$$Y_i^* = ((Y_{it_1} \otimes 1_c)^t, (Y_{it_1} \otimes 1_c)^t, \ldots, (Y_{it_2} \otimes 1_c)^t, \ldots, (Y_{it_{n_i-1}} \otimes 1_c)^t)^t,$$

$$Y_i^{**} = ((1_c \otimes Y_{it_2})^t, (1_c \otimes Y_{it_3})^t, \ldots, (1_c \otimes Y_{it_3})^t, \ldots, (1_c \otimes Y_{it_{n_i}})^t)^t, \quad t = 1, \ldots, n_i,$$

representam todas as combinações de pares distintos de respostas ordinais e 1_c são vetores de uns. Assim, os resíduos condicionais para equações de estimação de segunda ordem usando ALR é uma regressão de $\mathbf{Y}^*$ em $\mathbf{Y}^{**}$ de modo que a esperança condicional é dada por $\xi_i = E(Y^*|Y^{**})$.

2.2.3 Método GEE1 para dados ordinais

Segundo Heagerty e Zeger (1996), se o interesse primário está na estimação dos parâmetros β, assumindo a natureza ordinal da resposta, resulta nas equações escores representadas a seguir:

$$\begin{bmatrix} U_1(\beta,\alpha) \\ U_2(\beta,\alpha) \end{bmatrix} = \sum_{i=1}^{K} \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ 0 & \frac{\partial \sigma_i}{\partial \alpha} \end{bmatrix}^t \begin{bmatrix} V_{i11} & V_{i12}^c \\ V_{i21}^c & V_{i22}^c \end{bmatrix}^{-1} \begin{bmatrix} Y_i - \mu_i(\beta) \\ S_i - \sigma_i(\beta,\alpha) \end{bmatrix}, \tag{2.25}$$

em que, "c", representa as matrizes de covariância para os produtos $S_{i(t,s)} = (Y_{it} - \mu_{it}) \otimes (Y_{is} - \mu_{is})$ e $\sigma_i = E(S_i)$. Com isso, para o método GEE1, escreve-se separadamente as equações de estimação para β, e para os parâmetros de associação α :

$$U_1^*(\beta,\alpha) = \sum_{i=1}^{K} \left[\frac{\partial \mu_i}{\partial \beta}\right]^t {V_{i11}}^{-1}(Y_i - \mu_i(\beta))$$

e

$$U_2^*(\beta,\alpha) = \sum_{i=1}^{K} \left[\frac{\partial \sigma_i}{\partial \alpha}\right]^t \hat{V}_{i22}^{-1}(S_i - \sigma_i(\beta,\alpha)).$$

A estimação de $(\hat{\beta},\hat{\alpha})$ para dados ordinais seguem a mesma linha de estimação para dados binários, usando a log-razão de chances como uma medida da associação entre as respostas Y_{it} e Y_{il}.

Uma característica essencial para os dados ordinais é que a covariância do vetor de respostas para o $i-$ésimo grupo, Y_i, tem uma estrutura de bloco-diagonal de cada vetor de indicadores Y_{ij}, determinado pela média μ_{it}. Qualquer estrutura de associação de "trabalho", deverá preservar a estrutura de bloco diagonal da matriz peso, e para obtenção das estimativas $(\hat{\beta},\hat{\alpha})$, o procedimento iterativo é realizado separadamente, e assume inicialmente $\beta^{(0)}$, obtidas

sob suposição de independência de $\alpha^{(0)}$, ou seja, $\alpha^{(0)} = 0$ é dado por:

$$\beta^{(m+1)} = \beta^{(m)} + \left(\sum_{i=1}^{K} D_{i11}{}^t V_{i11}{}^{-1} D_{i11} \right)^{-1} \left(\sum_{i=1}^{K} U_1^*(\beta^{(m)}, \alpha^{(m)}) \right)$$

$$\alpha^{(m+1)} = \alpha^{(m)} + \left(\sum_{i=1}^{K} D_{i22}{}^t \hat{V}_{i22}^{-1} D_{i22} \right)^{-1} \left(\sum_{i=1}^{K} U_2^*(\beta^{(m)}, \alpha^{(m)}) \right)$$

em que $D_{i11} = \partial \mu_i / \partial \beta$, e $D_{i22} = \partial \sigma_i / \partial \alpha$.

2.2.4 Método GEE2 para dados ordinais

No caso da metodologia GEE2, os parâmetros da regressão do modelo marginal e de associação não são considerados independentes, isto é, o processo de estimação se dá de forma conjunta, e para obtenção das estimativas $(\hat{\beta}, \hat{\alpha})$, as matrizes de pesos V_{i22} e V_{i12} são formuladas para cada grupo, e as estimativas para os parâmetros (β, α) é solução das equações escores

$$S_\beta(\beta, \alpha) = \sum_{i}^{K} U_i(\beta, \alpha) = 0, \tag{2.26}$$

de modo que a contribuição do $i-$ésimo grupo para as equações (2.26) é dada por:

$$U_i(\beta, \alpha) = \begin{bmatrix} \frac{\partial \mu_i}{\partial \beta} & 0 \\ \frac{\partial \nu_i}{\partial \beta} & \frac{\partial \nu_i}{\partial \alpha} \end{bmatrix}^t \times \begin{bmatrix} V_{i11} & V_{i12} \\ V_{i21} & V_{i22} \end{bmatrix}^{-1} \times \begin{bmatrix} Y_i - \mu_i(\beta) \\ W_i - \nu_i(\beta, \alpha) \end{bmatrix},$$

em que

$$Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^t;$$

$$\mu_i = E(\mu_i);$$

$$W_i = \left[(Y_{i1} \otimes Y_{i2})^t, (Y_{i1} \otimes Y_{i3})^t, \ldots, (Y_{i(n_i-1)} \otimes Y_{in_i})^t \right]^t;$$

$$\nu_i = E(W_i); \quad V_{i11} = var(Y_i); \quad V_{i12} = cov(Y_i, W_i) \quad \text{e} \quad V_{i22} = var(W_i).$$

Agresti e Natarajan (2001) em uma revisão sobre as várias estratégias para modelar as variáveis de respostas categóricas ordinais, quando os dados ordenados possuem algum tipo de agrupamento, trataram em especial, as medidas repetidas que ocorrem em várias ocasiões como nos estudos longitudinais. Na formulação de modelos, as unidades de amostragem são os grupos. Em aplicações, cada grupo é um conjunto de medidas repetidas de um mesmo

indivíduo. Em outros, cada grupo é um conjunto de observações que se espera ser homogêneo. Assim, os modelos para as respostas ordinais, diferem em termos da média marginal, ou seja, a escolha do modelo afeta as interpretações dos parâmetros que descrevem a associação entre as respostas, e aos que estão relacionados ao modelo marginal.

Todavia, o método GEE2 conserva a propriedade de consistência, somente sob a suposição correta da especificação do modelo, porém possui a desvantagem de que as estimativas dos parâmetros do modelo marginal, $\hat{\beta}$, são não eficientes caso o modelo seja inapropriado (AGRESTI; NATARAJAN, 2001).

Especificado o modelo correto, o método GEE2 estima os parâmetros de associação mais precisamente e permite obter estimativas consistentes para os parâmetros do modelo marginal, realizando o procedimento iterativo para obtenção de α e β conjuntamente:

$$\begin{pmatrix} \beta^{(m+1)} \\ \alpha^{(m+1)} \end{pmatrix} = \begin{pmatrix} \beta^{(m)} \\ \alpha^{(m)} \end{pmatrix} + \left(\sum_{i=1}^{K} D_i{}^t V_i{}^{-1} D_i \right)^{-1} \left(\sum_{i=1}^{K} U_i(\beta^{(m)}, \alpha^{(m)}) \right), \tag{2.27}$$

em que $D_i = \partial(\mu_i, \nu_i)/\partial(\beta, \alpha)$ representa a matriz de derivadas para o $i-$ésimo grupo e $V_i = cov(Y_i, W_i)$.

2.2.5 Metodologia GEE usando a medida Kappa para dados ordinais

Em estudos da área médica, os ensaios clínicos, geralmente os pesquisadores estão interessados na avaliação que diferentes métodos ou procedimentos possam produzir valores semelhantes para medir variáveis de interesse. O coeficiente kappa, κ, ganhou popularidade nos estudos para avaliação de dois provadores em dois métodos, proposto por Lee, Koh e Ong (1989). Posteriormente, com a ampliação da abordagem de Lee, Koh e Ong (1989) para medidas repetidas(CHINCHILLI et al., 1996), o coeficiente κ tem sido amplamente utilizado nas áreas da saúde e biológicas.

O coeficiente Kappa, κ, é uma medida de concordância e apresenta valores entre -1 e 1, em que valores próximos de zero indicam que a concordância é a esperada pelo acaso e para valores próximos de 1 sugerem a não aleatorieadade das respostas. Para κ negativos, sugere que a concordância encontrada foi menor do aquela esperada pelo acaso, e portanto apontam discordância entre as respostas, porém seu valor não tem interpretação como intensidade de discordância. Kappa é baseado no número de respostas concordantes, ou seja, o número de

casos cujos resultados são os mesmos entre todos os indivíduos avaliados, e mede o grau de concordância além do que seria esperado somente pelo acaso é definida por:

$$k_{ist} = \frac{P_{oist} - P_{eist}}{1 - P_{eist}}, \tag{2.28}$$

em que P_{eist} é a probabilidade de que o par de variáveis categóricas sejam iguais assumindo independência e P_{oist} é a probabilidade conjunta dos pares de respostas serem iguais(COHEN, 1960).

Segundo, Klar, Lipsitz e Ibrahim (2000), Gonin et al. (2000), o ajuste de modelos para dados categóricos fornecem uma medida resumo, porém existe a necessidade prática de efetuar comparações entre grupos ou múltiplas amostras para determinar e avaliar a força de associação existente entre elas.

Dada a necessidade de obter informação sobre a concordância entre respostas correlacionadas, Williamson, Manatunga e Lipsitz (2000), Gonin et al. (2000) e Klar, Lipsitz e Ibrahim (2000) incorporaram covariáveis nas equações de estimação para modelagem dos pares de associações utilizando o coeficiente Kappa, κ, como medida de dependência longitudinal entre as respostas categóricas correlacionadas.

Ambos os autores construíram dois conjuntos de equações de estimação. O primeiro para modelar as distribuições marginais das respostas categóricas e o segundo conjunto de equações de estimação é introduzido para estimar κ, modelando variáveis binárias para descrever a concordância entre as respostas.

A abordagem do método GEE para modelar o coeficiente de correlação de concordância, κ, em Barnhart e Williamson (2001), tratou-se de um conjunto de três equações de estimação e torna-se viável na medida em que o interesse está relacionado na identificação de covariáveis, para formulação do modelo marginal e acomoda um teste para verificação da dependência das estimativas para o coeficiente kappa. A primeira equação de estimação refere-se às estimativas, $\hat{\beta}$. A segunda, está relacionada a obtenção das estimativas dos parâmetros da variância, σ^2, que será desnecessária se a estimativa de momentos para a variância for utilizada no terceiro conjunto de equações de estimação, no qual se concentra em obter as estimativas para o coeficiente de correlação de concordância.

Neste trabalho, o interesse está no estudo da modelagem kappa para medir a concordância das respostas para análise de dados categorizados semelhante ao proposto por Williamson, Manatunga e Lipsitz (2000).

Na formulação das equações de estimação utilizando o coeficiente κ, Williamson, Manatunga e Lipsitz (2000) consideraram K indivíduos avaliados em T_i tempos ou ocasiões diferentes, $i = 1,2,\ldots,K$.

A resposta de interesse é uma variável categórica, denotada por Z_{it}, assim $Z_{it} = k$ se a $t-$ésima resposta para o $i-$ésimo indivíduo for a categoria k, $k = 1,2,\ldots,c-1$. Dessa forma, o vetor de respostas, $\mathbf{Y}_i$, $T_i(c-1)\times 1$, consiste em variáveis aleatórias binárias, Y_{itk}, definidos da seguinte forma:

$$Y_{itk} = \begin{cases} 1, & \text{se} \quad Z_{it} = k \\ 0, & \text{caso contrário.} \end{cases} \tag{2.29}$$

Para a resposta ordinal, o modelo marginal de probabilidade acumuladas, ϑ_{itk}, associada a uma função de ligação, $g(\cdot)$, terá o vetor de probabilidades marginais denotada por $\pi_{itk} = P(Z_{it} = k) = P(Y_{it} = 1) = E(Y_{itk})$ de dimensão, $T_i(c-1)\times 1$, tais que $\vartheta_{itk} = P(Z_{it} \leq k), k = 1,2,\ldots,c-1$. E para o $i-$ésimo indivíduo, seja o conjunto de covariáveis $\mathbf{X}_i t$ em que o modelo marginal de parâmetros β, fica determinado como $g(\vartheta_{itk}) = X'_{itk}\beta$.

Consequentemente, para o primeiro conjunto de equações de estimação para a distribuição marginal das resposta é

$$v_1(\beta) = \sum_{i=1}^{K} \mathbf{D}_i^T \mathbf{V}_i^{-1}\{\mathbf{Y}_i - \pi_i\} = \mathbf{0}, \tag{2.30}$$

em que $\mathbf{D}_i = \partial\pi_i(\beta)/\partial\beta$, $\mathbf{V}_i = \mathbf{V}_i(\beta,\alpha) \approx var(\mathbf{Y}_i)$ é a matriz de covariância de trabalho de $\mathbf{Y}_i$(LIANG; ZEGER, 1986). Contudo, para a equação (2.29) sejam as respostas categóricas

$$P_{eist} = \sum_{i=1}^{K} \pi_{isk}\pi_{itk} \quad \text{e} \quad P_{oist} = \sum_{i=1}^{K} \omega_{istkk},$$

em que π_{isk} e π_{itk} são as probabilidades marginais do $i-$ésimo indivíduo ter como resposta a $k-$ésima categoria no $s-$ésimo e $t-$ésimo tempo de observação, e ω_{istkk} corresponde a probabilidade de que ambas as respostas do $i-$ésimo indivíduo sejam a $k-$ésima categoria.

Assim, para o segundo conjunto de equações de estimação, Williamson, Manatunga e Lipsitz (2000) construíram variáveis aleatórias binárias que descrevem a concordância entre as $s-$ésimas e $t-$ésimas respostas para o indivíduo i, seguindo as mesmas especificações em Liang, Zeger e Qaqish (1992) e Williamson, Kim e Lipsitz (1995), ou seja,

$$U_{ist} = \sum_{k=1}^{c} Y_{isk}Y_{itk} \quad \text{em que} \quad U_i = \{U_{i12}, U_{i13}, \ldots, U_{i,T_i-1,T_i}\}$$

Logo, o parâmetro de correlação de concordância, κ, é estimado resolvendo o segundo conjunto de equações de estimação:

$$\nu_2(\beta,\alpha)=\sum_{i=1}^{K}\mathbf{C}_i^T\mathbf{W}_i^{-1}\{\mathbf{U}_i-\mathbf{P}_{oi}(\alpha,\beta)\}=\mathbf{0}, \tag{2.31}$$

em que $\mathbf{W}_i$ é a matriz de covariância de trabalho de $\mathbf{U}_i$ de dimensão $T_i(T_i-1)/2\times T_i(T_i-1)/2$, e $\mathbf{C}=\partial\mathbf{P}_{oi}/\partial\alpha$. Segundo definido em Klar, Lipsitz e Ibrahim (2000) para valores de k_{ist} estimados pertencerem ao espaço paramétrico de κ, propuseram usar a inversa da transformação de Fisher's,

$$k_{ist}=\frac{exp(\mathbf{z'}_{ist}\alpha)-1}{exp(\mathbf{z'}_{ist}\alpha)+1} \tag{2.32}$$

em que $\mathbf{z'}_{ist}$ é um vetor de covariáveis do modelo κ, e k_{ist} é a medida de concordância entre a $s-$ésima e $t-$ésima resposta do $i-$ésimo indivíduo avaliado por κ.

De modo que, para computar as estimativas $(\hat{\beta},\hat{\alpha})$, realiza-se o processo iterativo semelhante às equações introduzidas por Williamson, Kim e Lipsitz (1995), a saber:

$$\hat{\beta}^{(m+1)}=\hat{\beta}^{(m)}-\left(\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i\right)^{-1}\left(\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}\{\mathbf{Y}_i-\pi_i(\hat{\beta}^{(m)})\}\right) \tag{2.33}$$

e

$$\hat{\alpha}^{(m+1)}=\hat{\alpha}^{(m)}-\left(\sum_{i=1}^{K}\mathbf{C}_i^T\mathbf{W}_i^{-1}\mathbf{C}_i\right)^{-1}\left(\sum_{i=1}^{K}\mathbf{C}_i^T\mathbf{W}_i^{-1}\{\mathbf{U}_i-P_{oi}(\hat{\beta}^{(m+1)},\hat{\alpha}^{(m)})\}\right) \tag{2.34}$$

2.3 Matriz de correlação de trabalho

Nessa seção serão apresentados os aspectos relevantes sobre a escolha da matriz de correlação de trabalho, bem como alguns critérios que serão utilizados neste trabalho. Tais critérios de seleção foram selecionados para serem comparados ao critério proposto na segunda parte que compõe esta tese, devido a semelhança na fomulação da motivação das suas construções.

2.3.1 A importância da especificação correta da matriz de correlação de trabalho

Conforme comentado na seção 2.2, para o caso em que $\mathbf{C}_i(\rho)$ é desconhecida, Liang e Zeger (1986), propuseram a extensão do uso de modelos lineares generalizados para dados

longitudinais baseados em quase-verossimilhança (NELDER; WEDDERBURN, 1972), cujas estimativas $\hat{\beta}$ é solução das equações (2.2).

Dessa forma, supondo que $\hat{\alpha}$ é um estimador consistente para α, Zhao, Prentice e Self (1992) avaliaram a eficiência de $\hat{\beta}_G$ (baseado na estrutura permutável ou AR(1)) com relação ao estimador $\hat{\beta}_T$, obtido sob suposição de verdadeira matriz de correlação para verificação da especificação incorreta da matriz de correlação de trabalho. Assim, $\hat{\beta}_T$ é denominada de verdadeira estimativa de quase-verossimilhança para o parâmetro de regressão.

Segundo Sutradhar e Das (2000) a avaliação computacional para comparação da eficiência das estimativas dos parâmetros de regressão deve ser realizada sob suposição dos valores limitantes das estimativas dos parâmetros de associação. E considerando que a eficiência computacional se baseia na matriz $\mathbf{R}(\alpha_0(\rho))$, em que $\alpha_0(\rho)$ é o valor limitante das estimativas dos parâmetros α, realizaram uma avaliação computacional com propósito de comparar a eficiência das estimativas dos parâmetros β com base em uma matriz de correlação de trabalho independente.

Posteriormente, efetuaram análise comparativa entre $\hat{\beta}_G$ e $\hat{\beta}_T$ (estimador de quase-verossimilhança de β assumindo a estrutura correta de correlação) sob $\mathbf{R}(\alpha_0(\rho))$ e confirmaram resultados apresentados em Sutradhar e Das (1999) de que os estimadores obtidos sob a suposição de independência produziram estimativas menos eficientes, quando comparado com $\hat{\beta}_G$ e além disso, que a eficiência das estimativas dos parâmetros β, depende da especificação da verdadeira estrutura de correlação de trabalho, e da magnitude dos parâmetros de correlação para obtenção das estimativas de α_0.

Existem três razões pelas quais uma escolha adequada da matriz de correlação de trabalho é importante, especialmente em termos de eficiência estatística.

A primeira razão, se dá pelo fato de que sob a má especificação da matriz de correlação de trabalho, a matriz de covariância sandwich:

$$\left(\sum_{i=1}^{K} D_i^T V_i^{-1} D_i\right)^{-1} \left(\sum_{i=1}^{K} D_i^T V_i^{-1} (Y_i - \mu_i)(Y_i - \mu_i)^T V_i^{-1} D_i\right) \left(\sum_{i=1}^{K} D_i^T V_i^{-1} D_i\right)^{-1} \tag{2.35}$$

é uma variância de propriedade assintótica e não pode ser assumida como válida em todas as situações. Pois, caso haja proporções maiores de dados faltantes, ou se o número de indivíduos for pequeno, mas o número de medidas para cada um dos indivíduos forem grandes, o estimador da variância sandwich não será recomendado, visto que para seu uso é implicitamente necessá-

rio que hajam grandes replicações independentes do vetor de respostas de modo que (LIANG; ZEGER, 1986):

$$\frac{1}{K}\sum_{i=1}^{K}(Y_i-\mu_i(\hat{\beta}))(Y_i-\mu_i(\hat{\beta}))^T \longrightarrow Cov(Y_i), \quad K\to\infty \tag{2.36}$$

A segunda razão é que uma estrutura de correlação de trabalho que se aproxime da verdadeira matriz de covariância produz estimativas mais eficientes, e a eficiência relativa assintótica depende também da disparidade entre a estrutura de correlação de trabalho e a verdadeira estrutura de covariância obtida após o ajuste (WANG; CAREY, 2003).

A terceira razão pela qual a escolha da matriz de correlação de trabalho é importante, se dá pela possibilidade de violação de uma das condições de regularidades propostas por Liang e Zeger (1986)(Teorema 2), ou seja, no processo iterativo para obtenção das estimativas $\hat{\beta}$, equação (2.3), as estimativas $\hat{\alpha}(\beta)$ convergem para algum valor limitante e $\mathbf{U}(\beta,\hat{\alpha})=\mathbf{0}$, deverá fornecer as estimativas assintóticas para o parâmetro β, e estes por sua vez são eficientes.

Algumas questões relativas aos parâmetros de associação $\hat{\alpha}$ foram abordadas por Crowder (1995), em que sob as suposições de uma estrutura de correlação de trabalho mal especificada, e se a verdadeira matriz de correlação fosse AR(1), mostraram que $\hat{\alpha}$ para a estrutura de correlação permutável não existe ou não tem solução única em certos casos.

No entanto, pode ser benéfico modelar cuidadosamente os parâmetros de correlação, pelos seguintes argumentos: (1) evitar a perda da eficiência na estimação dos parâmetros de regressão, que pode resultar da aplicação da estrutura de correlação de trabalho incorreta, em particular, para valores maiores da correlação e tamanhos moderados de amostras (ALBERT; MCSHANE, 1995); (2) devido a incerteza da definição da matriz de correlação de trabalho, a abordagem de Liang e Zeger (1986) pode, em alguns casos, levar a uma completa violação da estimativa dos parâmetros de regressão (SUTRADHAR; DAS, 2000); (3) evitar problemas com relação a inviabilidade na estimação dos parâmetros de correlação, que também podem resultar da má especificação da estrutura verdadeira (CROWDER, 1995; WANG; CAREY, 2003).

Em GEE, se a matriz de correlação de trabalho é corretamente especificada, então sob a hipótese do modelo de regressão correto, as estimativas $\hat{\beta}$ são assintoticamente ótimas e a matriz de variância estimada, $\hat{\mathbf{V}}_G$, para os parâmetros β, se reduz a (WANG; LIN, 2005):

$$\mathbf{V}_{opt}=\lim_{K\to\infty}[\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i^T]^{-1} \tag{2.37}$$

Além disso, a especificação da correta matriz de correlação de trabalho resulta em melhor eficiência.

A eficiência relativa assintótica para um estimador dos parâmetros de regressão é definido como o quociente entre os elementos das diagonais principais na referida matriz de covariância ótima, equação (2.37) e a da matriz de covariância estimada $\hat{\mathbf{V}}_G$.

2.3.2 Critérios de seleção da estrutura de correlação de trabalho

Dada a necessidade de obter estimativas consistentes e evitar a perda da eficiência na estimação dos parâmetros do modelo marginal sob abordagem GEE, um critério estatístico para seleção da estrutura de correlação de trabalho, deve ajudar na escolha razoável da matriz de correlação.

Rotnitzky e Jewell (1990) propuseram um teste estatístico para a hipótese de que o vetor dos coeficientes de regressão eram iguais a β, e que, se ambos os modelos marginais e a matriz de covariância para GEE fossem especificamente corretos, pode-se esperar que Ψ_0 e Ψ_1 são razoavelmente idênticos, em que se define respectivamente como segue:

$$\Psi_0 = K^{-1}\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i \quad , \tag{2.38}$$

$$\Psi_1 = K^{-1}\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}(\mathbf{Y}_i-\mu_i)(\mathbf{Y}_i-\mu_i)^T\mathbf{V}_i^{-1}\mathbf{D}_i \quad , \tag{2.39}$$

$$\Psi = \Psi_0^{-1}\Psi_1. \tag{2.40}$$

Quando a estrutura de correlação de trabalho é corretamente especificada, Ψ deverá estar próxima da matriz identidade. Hin, Carey e Wang (2007) descreveu o critério Rotnizky and Jewell's Criterion(RJ) para selecionar a estrutura de correlação de trabalho como:

$$RJ(\mathbf{R}) = \left[(1-tr(\Psi)/p)^2+(1-tr(\Psi^2)/p)^2\right]^{\frac{1}{2}}, \tag{2.41}$$

em que p é o número de covariáveis envolvidas no modelo e tr refere-se ao traço. Na literatura o critério é conhecido por "Rotnizky and Jewell's criterion (RJC)".

Posteriormente, Pan (2001) propôs uma abordagem sob a modificação do critério AIC para seleção de modelos na abordagem GEE e consequentemente como critério de seleção para estrutura de correlação de trabalho,sob a suposição de modelo independente para quase-verossimilhança(QIC).

Para seleção de modelos o critério AIC é bastante conhecido. Contudo, ele não pode ser utilizado para abordagem GEE, visto que é baseado em verossimilhança. Assim, Pan (2001), propôs um critério com base em quase-verossimilhança para auxiliar na escolha do melhor modelo ou estrutura de correlação cuja expressão é dada por:

$$QIC(\mathbf{R}) = -2\mathbf{Q}(\hat{\beta};\mathbf{I},\mathbf{D}) + 2tr(\hat{\Omega}\hat{\mathbf{V}}_G(\mathbf{R})), \tag{2.42}$$

em que $\hat{\mathbf{V}}_G(\mathbf{R})$ representa a matrix de covariância estimada a partir da estrutura de correlação de trabalho assumida, $\Omega = \sum_{i=1}^{K}(\mathbf{D}_i^T\mathbf{A}_i^{-1}\mathbf{D}_i|\mathbf{R})$ e, se a matriz de trabalho utilizada é a independente, $\mathbf{R} = \mathbf{I}$, sendo os pares de observação $(\mathbf{Y}_{it},\mathbf{X}_{it})$ em $\mathbf{D}$ independentes, então a quase-verossimilhança com base em $\mathbf{D}$ é:

$$\mathbf{Q}(\beta,\phi;\mathbf{I},\mathbf{D}) = \sum_{i=1}^{K}\sum_{t=1}^{n_i} Q(\beta,\phi,(\mathbf{Y}_{it},\mathbf{X}_{it})) \tag{2.43}$$

e assim, define-se o critério de seleção em Hardin (2005) quando $\Omega = \sum_{i=1}^{K}(\mathbf{D}_i^T\mathbf{A}_i^{-1}\mathbf{D}_i|\mathbf{I})$. Neste trabalho, a comparação dos resultados obtidos do critério proposto no segundo artigo, será com relação ao critério *QIC* formulado por Pan (2001). O critério é conhecido com "Quasi-likelihood under the independence model criterion (QIC)".

A Tabela 2.3 descreve algumas das funções de quase-verossimilhança comumente utilizada para as distribuições da família exponencial.

Tabela 2.3 – Funções de ligação e respectivas funções de quase-verossimilhança

Distribuição	Função de ligação	Função de quase-verossimilhança $Q(\beta,\phi,(\mathbf{Y}_{it},\mathbf{X}_{it}))$
Normal	μ_{it}	$(-1/2)(y_{it}-\mu_{it})^2$
Binomial	$ln\{\mu_{it}/(1-\mu_{it})\}$	$y_{it}ln\{\mu_{it}/(1-\mu_{it})\}+ln(1-\mu_{it})$
Poisson	$ln(\mu_{it})$	$y_{it}ln(\mu_{it})-\mu_{it}$
Gamma	$1/\mu_{it}$	$-y_{it}/\mu_{it}-ln(\mu_{it})$

Hin e Wang (2009) propuseram usar "metade" do segundo termo do critério QIC para selecionar a estrutura de correlação de trabalho na abordagem GEE, a estatística é chamanda de Critério de Informação de Correlação (CIC).

$$CIC = tr(\hat{\Omega}\hat{\mathbf{V}}_G(\mathbf{R})) \tag{2.44}$$

O primeiro termo do critério QIC, que se baseia em quase-verossimilhança, cujas funções de quase-verossimilhança, para algumas distribuições pertencentes a família exponencial está descrita na Tabela 2.3, está livre tanto da estrutura da matriz de correlação de trabalho como da verdadeira matriz de covariância.

Consequentemente, não fornece informações sobre a seleção da estrutura da matriz de covariância. Por outro lado, o segundo termo do critério QIC, contém informações sobre a estrutura de correlação através do estimador de variância de sandwich. Embora o segundo termo desempenhe um papel como uma penalização para a seleção de variáveis de modelo marginal, o critério QIC é mais "pesado" devido primeiro termo. Contudo, o critério QIC, não é uma medida particularmente sensível para seleção da estrutura de correlação de trabalho (HIN; WANG, 2009).

Gosho, Hamada e Yoshimura (2011) propuseram uma medida da discrepância entre o estimador da matriz de covariância e uma matriz de covariância especificada. Definiram que, como critério de seleção para a matriz de correlação de trabalho, a escolha entre as estruturas de matrizes avaliadas será àquela que minimiza $c(\mathbf{R})$, representada na equação:

$$c(\mathbf{R}) = tr\left[\left\{\left(\frac{1}{K}\sum_{i=1}^{K}(\mathbf{Y}_i-\mu_i)(\mathbf{Y}_i-\mu_i)^T\right)\left(\frac{1}{K}\sum_{i=1}^{K}\mathbf{V}_i\right)^{-1}-\mathbf{I}\right\}^2\right], \tag{2.45}$$

em que *tr* refere-se a soma dos elementos da digonal da matriz e $\mathbf{I}$ é a matriz identidade.

O critérios expostos nessa seção fazem uso somente das estimativas $\hat{\alpha}(\beta)$ na compoisção das matrizes de covariância estimadas. De modo que, em nenhum dos critérios mencionados são incorporadas as estimativas limitantes da matriz de covariância, sendo portanto fortemente influenciados pela magnitude das estimativas dos parâmetros de regressão.

Finalizando a metodologia base para obtenção e discussões dos resultados que serão apresentados na segunda parte deste trabalho, para as análises de simulação e ajustes de modelos, fez-se uso do Sistema Computacional Estatística **R** (R Core Team, 2015).

3 CONSIDERAÇÕES

Para validação das considerações presentes na pesquisa, estabelecendo relações que serão conceituadas na discussão dos resultados, e na perspectiva de proporcionar o embasamento teórico que fornece suporte ao desenvolvimento da segunda parte deste trabalho, a primeira parte constou do referencial base para a formulação das ideias descritas nos dois artigos que compõem o corpo desta tese.

Para tanto, concentrou-se nas equações de estimação generalizadas para dados ordinais, com a descrição dos modelos marginais e processos de estimação para os parâmetros de associação.

Posteriormente, dada a importância da incorporação de covariáveis nas equações de estimação generalizadas para a modelagem dos pares de associações, apresentou-se a metodologia GEE para dados ordinais utilizando o coeficiente Kappa, como medida da dependência longitudinal entre as respostas categóricas correlacionadas.

No que segue, escreveu-se a importância da escolha da matriz de correlação de trabalho para a metodologia GEE, e os aspectos sobre a eficiência e consistência dos parâmetros foram discutidos. Ao final, apresentou-se alguns dos critérios de seleção da literatura, bem como suas formulações.

REFERÊNCIAS

AGRESTI, A.; NATARAJAN, R. Modeling clustered ordered categorical data: A survey. **International Statistical Review**, Wiley Online Library, v. 69, n. 3, p. 345 – 371, 2001.

ALBERT, P. S.; MCSHANE, L. M. A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data. **Biometrics**, [Wiley, International Biometric Society], v. 51, n. 2, p. 627–638, 1995. ISSN 0006341X, 15410420. Disponível em: <http://www.jstor.org/stable/2532950>.

BARNHART, H. X.; WILLIAMSON, J. M. Modeling concordance correlation via gee to evaluate reproducibility. **Biometrics**, Wiley Online Library, v. 57, n. 3, p. 931–940, 2001.

BORÉM, F. M. Pós-colheita do café. **Lavras: UFLA**, v. 1, p. 631, 2008.

BORÉM, F. M. et al. Avaliação sensorial do café cereja descascado, armazenado sob atmosfera artificial e convencional. **Ciência e Agrotecnologia**, SciELO Brasil, v. 32, n. 6, p. 1724–1729, 2008.

CAREY, V.; ZEGER, S. L.; DIGGLE, P. Modelling multivariate binary data with alternating logistic regressions. **Biometrika**, Biometrika Trust, v. 80, n. 3, p. 517–526, 1993.

CHAGAS, S. J. de R.; MALTA, M. R.; PEREIRA, R. G. F. A. Potencial da região sul de minas gerais para a produção de cafés especiais (i–atividade da polifenoloxidase, condutividade elétrica e lixiviação de potássio). **Ciênc. agrotec.**, v. 29, n. 3, 2005.

CHINCHILLI, V. M. et al. A weighted concordance correlation coefficient for repeated measurement designs. **Biometrics**, JSTOR, p. 341–353, 1996.

COHEN, J. A coefficient of agreement for nomianal scales. **Educational Psychological Measurement**, v. 20, 1960.

CROWDER, M. On the use of a working correlation matrix in using generalised linear models for repeated measures. **Biometrika**, Biometrika Trust, v. 82, n. 2, p. 407–410, 1995.

GONIN, R. et al. Regression modelling of weighted κ by using generalized estimating equations. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 49, n. 1, p. 1–18, 2000.

GOSHO, M.; HAMADA, C.; YOSHIMURA, I. Criterion for the selection of a working correlation structure in the generalized estimating equation approach for longitudinal balanced data. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 40, n. 21, p. 3839–3856, 2011.

HARDIN, J. W. **Generalized estimating equations (GEE)**. [S.l.]: Wiley Online Library, 2005.

HEAGERTY, P. J.; ZEGER, S. L. Marginal regression models for clustered ordinal measurements. **Journal of the American Statistical Association**, Taylor & Francis, v. 91, n. 435, p. 1024–1036, 1996.

HIN, L.-Y.; CAREY, V. J.; WANG, Y.-G. Criteria for working correlation structure selection in gee. **The American Statistician**, v. 61, n. 4, p. 360–364, 2007.

HIN, L.-Y.; WANG, Y.-G. Working correlation structure identification in generalized estimating equations. **Statistics in medicine**, Wiley Online Library, v. 28, n. 4, p. 642–658, 2009.

ILLY, E. A saborosa complexidade do café. a ciência que está por trás de um dos prazeres simples da vida. **Revista Scientific American Brasil São Paulo**, n. 2, p. 48–53, 2002.

KLAR, N.; LIPSITZ, S. R.; IBRAHIM, J. G. An estimating equations approach for modelling kappa. **Biometrical Journal**, Wiley Online Library, v. 42, n. 1, p. 45–58, 2000.

LEE, J.; KOH, D.; ONG, C. Statistical evaluation of agreement between two methods for measuring a quantitative variable. **Computers in biology and medicine**, Elsevier, v. 19, n. 1, p. 61–70, 1989.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. **Biometrika**, Biometrika Trust, v. 73, n. 1, p. 13–22, 1986.

LIANG, K.-Y.; ZEGER, S. L.; QAQISH, B. Multivariate regression analyses for categorical data. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 3–40, 1992.

MALAVOLTA, E. **Historia do café no Brasil: Agronomia agricultura e Comercialização**. [S.l.]: Editora Agronômica Ceres Ltda., 2000.

MAZZAFERA, P. et al. Extração e dosagem da atividade da polifenoloxidase do café. **Scientia Agrícola**, São Paulo-Escola Superior de Agricultura"Luiz de Queiroz", 2002.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society A**, v. 135, p. 370–84, 1972.

PAN, W. Akaike's information criterion in generalized estimating equations. **Biometrics**, Wiley Online Library, v. 57, n. 1, p. 120–125, 2001.

PRENTICE, R. L.; ZHAO, L. P. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. **Biometrics**, JSTOR, p. 825–839, 1991.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2015. Disponível em: <https://www.R-project.org/>.

ROTNITZKY, A.; JEWELL, N. P. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. **Biometrika**, Biometrika Trust, v. 77, n. 3, p. 485–497, 1990.

SCAA. **Specialty Coffee Association of America. Cupping Specialty Coffee**. 2015. Http://scaa.org/?page=resources&d=cupping-protocols.

SCHMIDT, C. A. P.; MIGLIORANZA, É. A análise sensorial e o café: Uma revisão. **Revista Eletrônica Científica Inovação e Tecnologia**, v. 2, n. 2, p. 16–24, 2011.

SUTRADHAR, B. C.; DAS, K. Miscellanea. on the efficiency of regression estimators in generalised linear models for longitudinal data. **Biometrika**, Biometrika Trust, v. 86, n. 2, p. 459–465, 1999.

SUTRADHAR, B. C.; DAS, K. On the accuracy of efficiency of estimating equation approach. **Biometrics**, Wiley Online Library, v. 56, n. 2, p. 622–625, 2000.

VERBEKE, G. **Models for Discrete Longitudinal Data. Springer Series in Statistics**. [S.l.]: Springer, 2005.

WANG, Y.-G.; CAREY, V. Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. **Biometrika**, Biometrika Trust, v. 90, n. 1, p. 29–41, 2003.

WANG, Y.-G.; LIN, X. Effects of variance-function misspecification in analysis of longitudinal data. **Biometrics**, Wiley Online Library, v. 61, n. 2, p. 413–421, 2005.

WILLIAMSON, J. M.; KIM, K.; LIPSITZ, S. R. Analyzing bivariate ordinal data using a global odds ratio. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 90, n. 432, p. 1432–1437, 1995.

WILLIAMSON, J. M.; LIPSITZ, S. R.; MANATUNGA, A. K. Modeling kappa for measuring dependent categorical agreement data. **Biostatistics**, Biometrika Trust, v. 1, n. 2, p. 191–202, 2000.

ZHAO, L. P.; PRENTICE, R. L.; SELF, S. G. Multivariate mean parameter estimation by using a partly exponential model. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 805–811, 1992.

SEGUNDA PARTE - ARTIGOS

ARTIGO 1

Estratégia de modelagem via GEE em um experimento sensorial de cafés especiais caracterizados pela presença de diferentes grupos de múltiplas respostas ordinais repetidas

Artigo redigido conforme normas da Universidade Federal de Lavras
(Submetido à revista: Food Quality and Preference - Sujeito a alterações)

Estratégia de modelagem via GEE em um experimento sensorial de cafés especiais caracterizados pela presença de diferentes grupos de múltiplas respostas ordinais repetidas

RESUMO

A contribuição deste trabalho, mediante a aplicação proposta para avaliar a preferência e qualidade sensorial de genótipos de cafés especiais, é dada no aspecto metodológico diferenciando-se da abordagem usual de generalized estimating equation (GEE). Propõe-se incorporar múltiplas respostas ordinais repetidas, sendo essas caracterizadas pelas respostas categóricas repetidas por provadores e introduzir um terceiro conjunto de equações de estimação com o propósito de modelarmos as associações entre safras. Foram colhidas amostras de café (*Coffea arabica L.*) ao longo das safras de 2010/11, 2011/12, 2012/13 e 2013/14, o ambiente de cultivo do café foi estratificado em três classes de altitude e para cada um dos ambientes, foram coletados frutos amarelos representativos dos genótipos Bourbon Amarelo e Catuaí Amarelo e frutos vermelhos representativos dos genótipos Acaiá e Mundo Novo. Para todas as combinações envolvendo ambiente e genótipo, foram coletadas três repetições que foram avaliadas por quatro provadores e suas notas categorizadas. Concluiu-se que a estratégia proposta foi eficiente por discriminar as diferenças entre as categorias de notas mais elevadas e de menores notas, bem como a identificação dos atributos sensoriais que são semelhantes ao longo das safras.

Palavras-chave: Medidas repetidas. Análise sensorial. Categorias ordinais. Odds ratio. Coeficiente Kappa .

Modeling strategy with GEE in a sensory analysis of specialty coffees characterized by the presence of different groups of multiple repeated ordinal responses

ABSTRACT

The contribution of this work, in view of the proposed application to evaluate preference for and sensory quality of genotypes of specialty coffees, is given by a methodological aspect which differs from the usual approach of generalized estimating equation (GEE). We propose incorporating multiple repeated ordinal responses, which are characterized by the categorical repeated responses given by tasters, and introducing a third set of estimating equations to model the associations among crop seasons. Coffee samples (Coffea arabica L.) were collected along the crop seasons of $2010/11$, $2011/12$, $2012/13$ and $2013/14$. The coffee cultivation environment was divided into three altitude classes and, for each environment, yellow beans representing the Yellow Bourbon and Yellow Catuaí genotypes and red beans representing the Acaiá and Mundo Novo genotypes were collected. For all combinations involving environment and genotype, three replications were collected, which were evaluated by four tasters and their scores were categorized. It was concluded that the proposed strategy was efficient since it distinguishes the differences between the categories of higher and lower scores, as well as the identification of the sensory attributes which are similar throughout the crop seasons.

Keywords: Repeated measures. Sensory analysis. Ordinal categories. Odds ratio. Kappa coefficient.

1 INTRODUÇÃO

Na análise de dados com medidas repetidas, existe uma variedade considerável de técnicas quando a variável resposta segue uma distribuição normal: análise multivariada de perfis; análise de curvas de crescimento e modelos de regressão de efeitos aleatórios normais. Porém, não atentendo ao pressuposto de normalidade da variável resposta, uma série de dificuldades podem surgir devido à escassez de técnicas de análises que envolvam experimentos em análise sensoriais nos quais, as respostas podem ser de natureza categórica ordinal ou nominal.

Nesse contexto, um modelo que possibilite contemplar possíveis mudanças nas respostas dos indivíduos sob o tempo ou ocasições de observações, além de avaliar quais fatores influenciam a heterogeneidade entre indivíduos, torna-se viável o estudo de medidas correlacionadas entre as provas de xícaras, bem como a adaptação do parâmetro de associação na análise dos resultados experimentais. Com esse propósito se enquadram os modelos marginais obtidos por equações de estimação generalizadas(GEE) proposto por Liang e Zeger (1986).

Contudo, a abordagem GEE propõe analisar dados com medidas repetidas utilizando modelos lineares generalizados (NELDER; WEDDERBURN, 1972) e não pressupõe a especificação completa da distribuição multivariada das respostas repetidas. Logo, em se tratando de dados ordinais há o interesse na estimação dos parâmetros de associação representados pela razão de chances global como medida de associação obtida no ajuste dos modelos marginais para os pares de respostas repetidas ordinais.

A metodologia que se apresenta nesse artigo consiste em avaliar os resultados provenientes de análise sensorial da qualidade de cafés especiais, buscando associar possíveis mudanças dos atributos sensorias nas medidas repetidas obtidas pelas provas de xícaras. É uma proposta inovadora no sentido de possibilitar o uso da técnica em análise sensorial à produtos diversos nos quais possuem características de dados longitudinais ou simplesmente com múltiplas respostas repetidas. Por exemplo, pode-se está interessado no tempo e na identificação de covariáveis que influenciam a qualidade do produto final após o congelamento; ou o interesse pode está voltado às chances da aceitabilidade do produto; ou ainda, na determinação das probabilidades de classificação da qualidade do produto ao longo do tempo pelos consumidores.

Convém ressaltar que a utilização de GEE em análise sensorial, proporcionará a introdução de novas metodologias que permitam obter resultados mais apurados. Nesse sentido, exemplificamos a descrição de cafés especiais em que os provadores são independentes, porém os resultados provenientes de suas percepções sensoriais são correlacionadas. Tal questão, torna-

se mais complexa para uma análise estatística ao se considerar uma escala de notas discretas e ordinais, sendo portanto, uma alternativa promissora em relação aos métodos convencionais de análise.

Diferentemente da especificação da probabilidade conjunta em termos de parâmetros mistos que envolvem o primeiro e segundo momentos marginais e as demais ordens canônicas, Fitzmaurice e Laird (1993) e Zhao e Prentice (1990) desenvolveram estimadores de máxima verossimilhança considerando os momentos de ordem superiores como contrastes.

Posteriormente, decorrente do fato de que respostas categóricas em geral são correlacionadas, tendo por base medidas longitudinais, o uso da abordagem GEE as respostas categóricas ordinais proposta por Heagerty e Zeger (1996) tem sido aprimorado e aplicado em diversas áreas do conhecimento. Liang e Zeger (1986); Clayton (1992); Gange et al. (1993); Williamson e Kim (1996), propuseram técnicas de regressão para modelagem de dados longitudinais para resposta multinomial, no qual desenvolveram técnicas de regressão para médias marginais utilizando global odds ratios como medida de associação em estudos oftalmológicos, porém não há registros do uso de tal metodologia em análise sensorial, bem como quando se consideram as associações entre as degustações realizadas e entre safras.

Williamson, Manatunga e Lipsitz (2000), Gonin et al. (2000) e Klar, Lipsitz e Ibrahim (2000) incorporaram covariáveis nas equações de estimação para modelagem dos pares de associações utilizando o coeficiente Kappa, κ, como medida de dependência longitudinal entre as respostas categóricas correlacionadas. Ambos os autores construíram dois conjuntos de equações de estimação. O primeiro para modelar as distribuições marginais das respostas categóricas e o segundo conjunto de equações de estimação é introduzido para estimar κ, modelando variáveis binárias para descrever a concordância entre as respostas.

Convém ressaltar que em nenhum dos estudos anteriormente citados constam da associação entre as respostas repetidas longitudinais, avaliadas dentro de um mesmo grupo e obtidas separadamente, como elementos para aproximação da matriz de covariâncias utilizando GEE. Ou seja, tais associações por grupo, não foram inseridas em um modelo mais amplo para obtenção das respostas marginais. Em particular, não há registros de aplicações envolvendo produtos alimentícios cujas avaliações sensoriais, dadas em escalas ordinais, envolvam as associações entre degustações e entre as safras.

O presente artigo dá ao pesquisador em análise sensorial, ou aqueles cujo interesse está sobre dados do tipo categóricos ordinais, a oportunidade de realizar análises estatísticas sob

poucas suposições, além de apresentar uma metodologia já consagrada na literatura que possibilitará estudos em análises sensoriais sob novas vertentes.

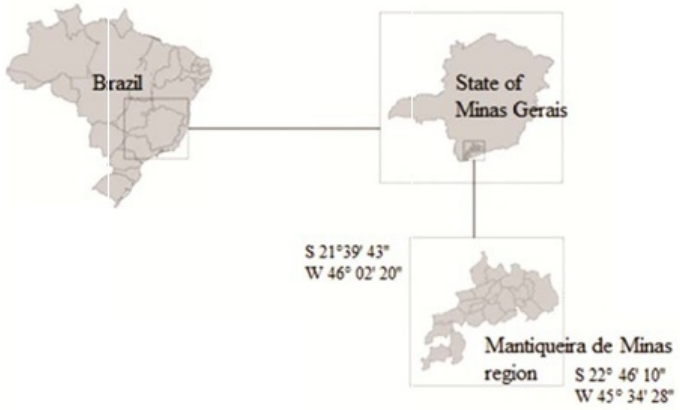
Com essa motivação, o presente trabalho tem por objetivo propor uma estratégia de modelagem de dados categóricos ordinais em um experimento sensorial de diferentes genótipos de cafés especiais, em que considera as associações entre as respostas múltiplas repetidas de um mesmo grupo de indivíduos para distintas safras, obtidas de diferentes subgrupos correlacionados.

2 METODOLOGIA

2.1 Descrição da área experimental e atributos sensoriais utilizados

A região de estudo possui extensão territorial de $6.317,38\ km^2$ com altitudes variando de 800 a 2.300 *m* acima do nível do mar. Após o levantamento de campo e reconhecimento da microrregião e considerando a grande extensão de abrangência do projeto e a complexidade da paisagem da Mantiqueira de Minas, optou-se por selecionar uma área piloto para a coleta das amostras de café, conforme Figura 1.

Figura 1 – Localização da região da Serra da Mantiqueira, estado de Minas Gerais, Brasil. Fonte: Ramos *et al*, 2016.



Para o presente estudo, foram coletadas amostras de café (*Coffea arabica L.*), ao longo de quatro safras $(2010/11,\ 2011/12,\ 2012/13$ e $2013/14)$, em lavouras comerciais de propriedades localizadas no município de Carmo de Minas, Minas Gerais, Brasil. O delineamento experimental foi baseado no estudo da interação entre variáveis ambientais, genéticas e de processamento.

O ambiente de cultivo do café foi estratificado em três classes de altitude (inferior a 1.000 *m*, entre 1.000 e 1.200 *m* e superior a 1.200 *m*) e dois grupos de vertentes, Sol (NE, N, NO e O) e Sombra (L, SE, S e SO), resultando na combinação de seis variáveis ambientais. Para cada um dos ambientes, foram coletados frutos amarelos representativos dos genótipos Bourbon Amarelo e Catuaí Amarelo e frutos vermelhos representativos dos genótipos Acaiá e Mundo Novo. Para todas as combinações envolvendo ambiente e genótipo, foram coletadas três repetições e processadas nas duas formas distintas (Via seca e Úmida), totalizando 72 amostras por safra.

A análise sensorial foi realizada por quatro provadores treinados e qualificados como juízes certificados de cafés especiais, utilizando-se a metodologia proposta pela Associação Americana de Cafés Especiais - SCAA (LINGLE, 2011). Em cada avaliação, foram degustadas cinco xícaras de café representativas de cada amostra. Nessa avaliação, foram atribuídas notas no intervalo de 0 a 10 pontos para cada um dos seguintes atributos: fragrância/aroma, uniformidade, ausência de defeitos, doçura, sabor, acidez, corpo, finalização, equilíbrio e impressão global. O conjunto de dados em análise sensorial de cafés especiais foram obtidos da realização do projeto "Protocolo de identidade, qualidade e rastreabilidade para embasamento da indicação geográfica dos cafés da Mantiqueira" aprovado no edital CNPq/MAPA 064/2007 (BOREM, 2007).

2.2 Especificações para construção do modelos

Para efeito de simplificação, denota-se por $1, 2, 3$ e 4 as respectivas safras avaliadas. Conforme anteriormente citado, há quatro genótipos de cafés que foram degustados, porém o número de genótipos avaliados para cada provador em cada safra não foi o mesmo, a saber: todos os provadores degustaram 4 variedades de cafés na safra 1; na safra 2, degustaram 3 variedades; nas safras 3 e 4, duas variedades de cafés.

Dessa forma, para cada uma das 72 amostras por safra, quatro provadores forneceram, cada um, 288 avaliações sensoriais por safra, totalizando 288×4 amostras de cafés degustadas, de modo que, para safra 1, obteve-se 16 grupos de tamanhos, $30, 32, 4$ e 6, para safra 2, 12 grupos de tamanhos $36, 35$ e 1, e para as safras 3 e 4, 8 grupos de tamanhos 36. A Tabela 1 descreve a formação dos 44 grupos.

Tabela 1 – Contagens das notas finais por provadores, safra e genótipos

Grupos			Notas distribuídas por categorias		
Provador	Safra	Genótipo	1°Categoria < 82	2°Categoria $[82-91]$	3°Categoria > 91
1	1	1	15	14	1
		2	5	22	5
		3	1	2	1
		4	2	4	0
	2	1	19	16	1
		2	5	23	7
		3	0	0	1
	3	1	9	26	1
		2	0	33	3
	4	1	17	18	1
		2	16	30	0
2	1	1	12	16	2
		2	13	16	3
		3	2	1	1
		4	2	4	0
	2	1	21	15	0
		2	7	26	2
		3	0	1	0
	3	1	9	27	0
		2	1	33	2
	4	1	15	21	0
		2	1	35	0
3	1	1	11	18	1
		2	3	25	4
		3	0	3	1
		4	1	5	0
	2	1	16	20	0
		2	4	24	7
		3	0	1	0
	3	1	1	34	1
		2	1	29	6
	4	1	8	28	0
		2	3	33	0
4	1	1	9	16	5
		2	2	24	6
		3	0	2	2
		4	1	5	0
	2	1	14	20	2
		2	7	15	13
		3	0	0	1
	3	1	8	26	2
		2	1	27	8
	4	1	7	29	0
		2	2	33	1

Portanto, onde consta $genótipo_j$ entende-se por um grupo de variedades de cafés degustadas na $j-$ésima safra e de maneira análoga, entende-se o conjunto de todos os provadores da $j-$ésima safra por $provador_j$. Desse modo, $\mathbf{O}_{ij} = \{O_{ij_1}, O_{ij_2}, \ldots, O_{ij_{n_i}}\}$ representa o vetor de avaliações dadas pelo $i-$ésimo provador na $j-$ésima safra.

Para efeito de aplicação da metodologia, considerou-se a resposta de interesse como a classificação das notas dadas aos café especiais nas categorias (1 : notas finais < 82 ; 2 : notas finais $82-91$ inclusive; 3 : notas finais > 91), avaliadas para o $i-$ésimo provador na $j-$ésima safra, $i,\ j = 1,2,3,4$.

2.3 Procedimentos de organização e estruturação dos dados categóricos

O estudo foi dividido em dois procedimentos: primeiro, conforme Tabela 2, considerou-se somente os percentuais relacionados aos grupos de provadores e safras, realizando a categorização das notas finais, totalizando 16 amostras referente às contagens por categorias.

Tabela 2 – Contagens e percentuais das notas finais por categorias segundo provador e safra

Variáveis		Notas distribuídas por categorias		
Provadores	Safras	1° Categoria $(<82)(\%)$	2° Categoria $[82-91](\%)$	3° Categoria $(>91)(\%)$
1	1	23(28,75)	42(22,34)	7(33,33)
	2	24(30,0)	39(20,74)	9(42,86)
	3	9(11,25)	59(31,38)	4(19,05)
	4	24(30,0)	48(24,53)	1(4,76)
2	1	29(34,94)	37(18,97)	6(60)
	2	28(33,73)	42(21,54)	2(20)
	3	10(12,05)	60(30,77)	2(20)
	4	16(19,28)	56(28,72)	0(0)
3	1	15(31,25)	51(23,18)	6(30)
	2	20(41,67)	45(20,45)	7(35)
	3	2(4,17)	63(28,64)	7(35)
	4	11(22,92)	61(27,73)	0(0)
4	1	12(23,53)	47(23,86)	13(6,60)
	2	21(41,18)	35(17,77)	16(8,12)
	3	9(17,65)	53(26,90)	10(5,08)
	4	9(17,65)	62(31,47)	1(0,51)

Os percentuais de interesse são os que indicam uma associação nas mudanças que ocorrem nas notas dentro das categorias por safra, visando assim determinar um possível efeito de safra para a classificação das notas, ou ainda, detectar a preferência dos provadores por uma ou outra categoria.

Posteriormente, ainda nesse procedimento, considerando que os genótipos de cafés são determinantes para a classificação das notas finais, realizou-se as contagens das notas distribuídas por categorias, conforme Tabela 1, ajustou-se modelos logito para categorias adjacentes na perspectiva de determinarmos probabilidades para a preferências das notas sob as categorias associadas às safras e provadores.

Para o segundo procedimento, utilizando as covariáveis altitude, processamento e vertente, ajustou-se os modelos marginais para os 44 grupos distintos obtidos da combinação entre os fatores definidos por provadores, $i = 1,\ldots,4$, safra, $j = 1,\ldots,4$ e quantidade de genótipos avaliados em cada safra (ver Tabela 1), seguindo a estratégia de modelagem para construção do modelo marginal geral proposto neste artigo.

2.4 Modelo logito para Categorias de notas adjacentes

Dado que as notas categorizadas obtidas de cada provador em cada uma das 72 amostras por safra possuem uma ordenação natural, segundo (AGRESTI, 2013) os logits ordinais podem ser usados como pares de probabilidades de respostas adjacentes.

Contextualizando para o conjunto de dados estudo nesse artigo, os logits de categorias adjacentes para k categorias de respostas, são definidos como:

$$logit[P(\text{Notas} = k|\text{Notas} = k \quad \text{ou} \quad k+1)] = log\frac{\pi_k}{\pi_{k+1}}, \quad k = 1,\ldots c-1. \tag{2}$$

E sendo a razão de chances propocionais, o modelo logit de categorias adjacentes fica determinado por

$$log\frac{\pi_k(\mathbf{x})}{\pi_{k+1}(\mathbf{x})} = \theta_k + \mathbf{x}^{\mathbf{T}}\beta, \quad k = 1,\ldots,c-1, \tag{3}$$

com efeitos β comuns para cada dos $c-1$ modelos logit.

2.5 Construção dos Modelos Marginais

Diante da proposta desse artigo, denotando $\mathbf{O}_i = \{\mathbf{O}^t_{i1}, \mathbf{O}^t_{i2}, \mathbf{O}^t_{i3}, \mathbf{O}^t_{i4}\}^t$ como um vetor de medidas ordinais para o $i-$ésimo provador, O_{ij_l}, $l = 1,2,\ldots,n_i$, representa a j_l-ésima

observação para o $i-$ésimo provador. A medida ordinal $O_{ij_l} = k$, $k = 1,2,3\ldots,c$, com c categorias de respostas corresponde a um vetor de variáveis indicadoras acumuladas $Y_{ij_l(k)} = I_{(O_{ij_l}>k)}$, $k \in 1,2,\ldots,c-1$ tal que

$$Y_{ij_l(k)} = \begin{cases} 1, & \text{se } O_{ij_l} > k \\ 0, & \text{caso contrário} \end{cases} \tag{4}$$

em que o modelo de razão de chances proporcionais para médias marginais é dado por:

$$logit[E(Y_{ij(k)})] = \theta_k + \mathbf{x}_{ij}^T\beta. \tag{5}$$

Para cada resposta O_{ij_l} associou-se um vetor $\mathbf{x}$ de p covariáveis $\mathbf{x}_{pj_l}$, de modo que, fixado o $i-$ésimo provador, $\mathbf{x} = (x_1, x_2, \ldots, x_p)^t$ indica o vetor de covariáveis observadas em cada tempo de degustação, $l = 1,2,\ldots,n_i$ na $j-$ésima safra. Na Tabela 3, segue uma representação do layout para os dados desse artigo.

Tabela 3 – Estrutura de um conjunto de dados longitudinais com p covariáveis associadas às respostas $\mathbf{O}_i$ para n provadores em l tempos de observação na $j-$ésima safra.

Provador (i)	Covariávies (x)				Observações	Respostas				Vetor de respostas
	1	2	...	p		O_{i1}	O_{i2}	...	O_{ij}	
1	x_{11}	x_{12}	...	x_{1p}	1	O_{11_1}	O_{12_1}	...	O_{1j_1}	
1	x_{21}	x_{22}	...	x_{2p}	2	O_{11_2}	O_{12_2}	...	O_{1j_2}	O_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
1	$x_{n_1 1}$	$x_{n_1 2}$	...	$x_{n_1 p}$	n_1	$O_{11_{n_1}}$	$O_{12_{n_1}}$	...	$O_{1j_{n_1}}$	
2	x_{11}	x_{12}	...	x_{1p}	1	O_{21_1}	O_{22_1}	...	O_{2j_1}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	O_2
2	$x_{n_1 1}$	$x_{n_1 2}$	...	$x_{n_1 p}$	n_2	$O_{21_{n_2}}$	$O_{22_{n_2}}$	...	$O_{2j_{n_2}}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{11}	x_{12}	...	x_{1p}	1	O_{n1_1}	O_{n2_1}	$\vdots$	O_{nj_1}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	O_n
n	$x_{n_n 1}$	$x_{n_n 2}$	...	$x_{n_n p}$	n_n	$O_{n1_{n_n}}$	$O_{n2_{n_n}}$	...	$O_{nj_{n_n}}$	

Desse modo, o vetor de respostas para $i-$ésimo provador na j_l-ésima observação, $\mathbf{Y}_{ij_l}^t$ segue uma distribuição Bernoulli com média $\mu_{ij_l} = P(Y_{ij_l} = 1)$. Logo, o vetor de respostas binárias para o $i-$ésimo provador é dado por $\mathbf{Y}_i = \{\mathbf{Y}_{i1}^t, \mathbf{Y}_{i2}^t, \ldots, \mathbf{Y}_{ij}^t\}^t$ em que, $\mathbf{Y}_{i1} = (\mathbf{Y}_{i1_1}^t, \mathbf{Y}_{i1_2}^t, \ldots, \mathbf{Y}_{i1_{n_i}}^t)^t$, $\mathbf{Y}_{i2} = (\mathbf{Y}_{i2_1}^t, \mathbf{Y}_{i2_2}^t, \ldots, \mathbf{Y}_{i2_{n_i}}^t)^t$, $\ldots$, $\mathbf{Y}_{ij} = (\mathbf{Y}_{ij_1}^t, \mathbf{Y}_{ij_2}^t, \ldots, \mathbf{Y}_{ij_{n_i}}^t)^t$, $j = 1,2,\ldots,n$ e $\mu_i = E(\mathbf{Y}_i)$.

Note que para a categoria de respostas k, $(k = 1,2,\ldots,c)$ associadas às covariáveis sensoriais, $\mathbf{x}_i = (\mathbf{x}_{1j}, \mathbf{x}_{2j}, \ldots, \mathbf{x}_{pj})$, o vetor de respostas para a $k-$ésima categoria, $\mathbf{Y}_{ij_l(k)}^t$, terá distri-

buição binomial com probalidade de sucesso $\pi_k(\mathbf{x_i})$. Para compreensão da estruturas das respostas, $\mathbf{Y}_{ij_l(k)}$ segue a Tabela 4, em que as notas dadas aos cafés especiais pelos quatro provadores, colhidas em quatro safras foram distribuídas em três categorias de notas ($k = 1,2,3.$)

Tabela 4 – Exemplo de respostas Y_i para 3 categorias de notas para 4 provadores em 4 safras com l observações.

Provador (i)	Observação (l)	Safra (j)	Resposta ordinais O_{ij_l}	Indicadoras $(k=1,2)$	Valores de Y_{ij_l}	Respostas Y_i
1	1	1	3	$\left(I_{(3>1)}, I_{(3>2)}\right)$	$(1,1)$	$(1,1)^t$
1	2	1	2	$\left(I_{(2>1)}, I_{(2>2)}\right)$	$(1,0)$	$(1,0)^t$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	n_1	1	2	$\left(I_{(2>1)}, I_{(2>2)}\right)$	$(1,0)$	$(1,0)^t$
1	1	2	1	$\left(I_{(1>1)}, I_{(1>2)}\right)$	$(0,0)$	$(0,0)^t$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	3	3	$\left(I_{(3>1)}, I_{(3>2)}\right)$	$(1,1)$	$(1,1)^t$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	4	1	$\left(I_{(1>1)}, I_{(1>2)}\right)$	$(0,0)$	$(0,0)^t$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	n_1	4	3	$\left(I_{(3>1)}, I_{(3>2)}\right)$	$(1,1)$	$(1,1)^t$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	n_4	4	2	$\left(I_{(2>1)}, I_{(2>2)}\right)$	$(1,0)$	$(1,0)^t$

Considerando três categorias de respostas, $(k = k_1, k_2, k_3)$, a razão de chances para o par O_{ij_h} e O_{ij_s} conforme descrito na Tabela 3, é um modelo de razão de chances proporcional que pode ser visualizado como uma regressão logística conjunta para cada uma das possíveis respostas binárias $Y_{ij_l(k)}$, definido como:

$$\psi_{i(j_h,j_s)(k_1,k_2)} = \frac{P(O_{ij_h} > k_1, O_{ij_s} > k_2)P(O_{ij_h} \leq k_1, O_{ij_s} \leq k_2)}{P(O_{ij_h} > k_1, O_{ij_s} \leq k_2)P(O_{ij_h} \leq k_1, O_{ij_s} > k_2)}. \tag{6}$$

Dessa forma, o número de possibilidades de interações para cada par (O_{ij_h}, O_{ij_s}) correspondente as categorias de respostas $k = 1,2,\ldots,c-1$ fixada, são de $(c-1)^2$ pares de razão de chances. Portanto, para n grupos de provadores o número de parâmetros ψ por safra será de $n_\psi = \sum_{i=1}^{n} \binom{n_i}{2} (c-1)^2$.

Note que na Tabela 4, $n_i (i = 1,2,3,4)$ se refere ao número de degustações realizadas por cada um dos provadores em cada uma das safras. O interesse se concentrou nos pares de razão de chances para as categorias de notas k_1 e k_2, a fim de identificarmos possíveis mudanças que caracterizam efeitos de safra. O modelo permutável para todas as j_h e j_s-ésimas degustações

foi considerado comum a todos os provadores, ou seja:

$$log(\psi_{i(j_h,j_s)(k_1,k_2)}) = \alpha, \quad \forall i = 1,2,3,4. \tag{7}$$

Conforme definido em (4) e utilizando $\psi_{i(j_h,j_s)}$ como medida de associação das respostas dentro de um mesmo grupo (provador, safra e genótipo), a expressão para razão de chances global (todas os pares de combinações possíveis de razões de chances) para $Y_{ij_h(k_1)}$ e $Y_{ij_s(k_2)}$ nomeada por $OR(Y_{ij_h(k_1)}, Y_{ij_s(k_2)})$ é estimada por:

$$logOR(Y_{ij_h(k_1)}, Y_{ij_s(k_2)}) = log\left(\frac{P(Y_{ij_h}=1, Y_{ij_s}=1)P(Y_{ij_h}=0, Y_{ij_s}=0)}{P(Y_{ij_h}=1, Y_{ij_s}=0)P(Y_{ij_h}=0, Y_{ij_s}=1)}\right). \tag{8}$$

Dessa forma, para a especificação do modelo marginal proposto por Heagerty e Zeger (1996), a correlação entre as respostas para modelos de regressão de razão de chances definida por

$$\rho_{i(j_h,j_s)(k_1,k_2)(\alpha)} = Corr(Y_{ij_hk_1}, Y_{ij_sk_2}|\mathbf{X_{ijhjs}}) = \frac{exp(\mathbf{X}_{ij_hj_s}\alpha)-1}{exp(\mathbf{X}_{ij_hj_s}\alpha)+1}$$

é a correlação para as variávies definidas em (4), e é obtida em função do vetor de parâmetros α, na qual a estrutura de correlação para as múltiplas respostas depende de covariávies sensoriais $\mathbf{X}_{ij_hj_s}$ através da função de ligação $g(\rho_{i(j_hj_s)}) = \mathbf{X}_{ij_hj_s}\alpha$ pelo seguinte modelo linear generalizado

$$log\left(\frac{1+\rho_{i(j_h,j_s)(k_1,k_2)}}{1-\rho_{i(j_h,j_s)(k_1,k_2)}}\right) = \mathbf{z}^t_{i(j_h,j_s)(k_1,k_2)}\alpha, \quad 1 \leq h < s \leq n_i, \quad i,j = 1,2,3,4, \tag{9}$$

em que $\mathbf{z}$ é um subconjunto de $(\mathbf{x}_{ij_h}, \mathbf{x}_{ij_s})$ ou qualquer outra covariável relevante para modelar o grau de associação entre as j_h e j_s-ésima degustações, inclusive pode-se assumir o modelo permutável como na equação (7), ou seja, $\rho_i = Corr(Y_{ij_h}, Y_{ij_s})$, para todo $h \neq s$.

Dessa forma, as expressões (8) e (9) são dadas para quantificar a associação entre as observações j_h e j_s em relação ao $i-$ésimo provador para cada uma das safras como:

$$logOR(\mathbf{Y}_{ij_h}, \mathbf{Y}_{ij_s}) = log\left(\frac{1+\rho_{i(j_h,j_s)}}{1-\rho_{i(j_hj_s)}}\right) = \mathbf{X}_{ij_hj_s}\alpha, \quad 1 \leq h < s \leq n_i, \quad i,j = 1,2,3,4. \tag{10}$$

Seguindo a estimação para medidas ordinais usando razão de chances global como medida de associação, duas equações de estimação, uma para obtenção dos parâmetros β e outra para os parâmetros α foram propostas conforme segue respectivamente (HEAGERTY; ZEGER, 1996) :

$$\mathbf{U}_1^*(\beta,\alpha) = \sum_{i=1}^{K} \left[\frac{\partial \mu_i}{\partial \beta}\right]^t \mathbf{V}_{i11}^{-1}(\mathbf{Y_i} - \mu_{\mathbf{i}}(\beta)) = \mathbf{0} \tag{11}$$

e

$$\mathbf{U}_2^*(\beta,\alpha) = \sum_{i=1}^{K} \left[\frac{\partial \sigma_i}{\partial \alpha}\right]^t \mathbf{V}_{i22}^{-1}(\mathbf{S_i} - \sigma_{\mathbf{i}}(\beta,\alpha)) = \mathbf{0} \tag{12}$$

em que $\mathbf{S}_{i(j_h,j_s)} = (\mathbf{Y}_{ij_h} - \mu_{\mathbf{ij_h}}) \otimes (\mathbf{Y}_{ij_s} - \mu_{\mathbf{ij_s}})$ e $\sigma_i = E(\mathbf{S}_i)$, $\mathbf{V}_{i11} = var(\mathbf{Y}_i), \mathbf{V}_{i22} = var(\mathbf{W}_i)$, sendo $\mathbf{W}_i$ formado por todas as combinações de pares distintos de respostas ordinais, isto é

$$\mathbf{W}_i = ((Y_{ij_1} \otimes Y_{ij_2})^t, (Y_{ij_1} \otimes Y_{ij_3})^t, \ldots, (Y_{ij_2} \otimes Y_{ij_3})^t, \ldots, (Y_{ij_{n_i-1}} \otimes Y_{ij_{n_i}})^t)^t, \quad j = 1,2,3,4,$$

representando cada resposta ordinal através do vetor $\mathbf{Y}_{ij_h}$ em que considera os K^2 produtos binários $\mathbf{Y}_{ij_h} \otimes \mathbf{Y}_{ij_s}$, para todo $h < s$.

Para computar $(\hat{\beta}, \hat{\alpha})$, usou-se Fisher-scoring-type algorithm tal que

$$\beta^{(m+1)} = \beta^{(m)} - \left(\sum_{i=1}^{K} \mathbf{D}_{i11}{}^t \mathbf{V}_{i11}{}^{-1} \mathbf{D}_{i11}\right)^{-1} \left(\sum_{i=1}^{K} \mathbf{U}_1^*(\beta^{(m)}, \alpha^{(m)})\right)$$

$$\alpha^{(m+1)} = \alpha^{(m)} - \left(\sum_{i=1}^{K} \mathbf{D}_{i22}{}^t \hat{\mathbf{V}}_{i22}^{-1} \mathbf{D}_{i22}\right)^{-1} \left(\sum_{i=1}^{K} \mathbf{U}_2^*(\beta^{(m)}, \alpha^{(m)})\right)$$

em que $\mathbf{D}_{i11} = \partial \mu_i / \partial \beta$, e $\mathbf{D}_{i22} = \partial \sigma_i / \partial \alpha$, sendo $m = 0, 1, \ldots$ o número de iterações.

2.6 Modelando as associações entre safras com coeficiente Kappa

Para a construção do conjunto de equações de estimação para as associações entre safras, sejam s e t, $(s < t)$ os pares de respostas para a $i-$ésima degustação. O coeficiente Kappa, κ, é uma medida de concordância das avaliações sensoriais e apresenta valores entre -1 e 1, em que valores próximos de zero indicam que a concordância é a esperada pelo acaso e para valores próximos de 1 sugerem a não aleatorieadade das respostas. Para κ negativos, sugere

que a concordância encontrada foi menor do aquela esperada pelo acaso e portanto, apontam discordância entre as respostas dos avaliadores, porém seu valor não tem interpretação como intensidade de discordância. Kappa é baseado no número de respostas concordantes, ou seja, o número de casos cujos resultados são os mesmos entre todos os avaliadores e mede o grau de concordância além do que seria esperado somente pelo acaso é definida por:

$$k_{ist} = \frac{P_{oist} - P_{eist}}{1 - P_{eist}}, \tag{13}$$

em que P_{eist} é a probabilidade de que o par de variáveis categóricas sejam iguais, assumindo independência, e P_{oist} é a probabilidade conjunta dos pares de respostas serem iguais. Mediante a nossa proposta, definimos $P_{eist} = P(Y_{is(k_1)} = 1)P(Y_{it(k1)} = 1) + P(Y_{is(k_2)} = 1)P(Y_{it(k_2)} = 1)$ e $P_{oist} = P(Y_{is(k1)} = 1, Y_{it(k_1)} = 1) + P(Y_{is(k_2)} = 1, Y_{it(k_2)} = 1)$, $s,t = 1,2,3,4, s < t$.

Seguindo o sugerido por Williamson, Manatunga e Lipsitz (2000), o produto das variáveis indicadoras descrevem a concordância entre as respostas da $s-$ésima e $t-$ésima safras, ou seja, $U_{ist} = Y_{is(k_1)}Y_{it(k_1)} + Y_{is(k_2)}Y_{it(k_2)}$ tais que $\mathbf{P}_{oi}(\alpha^*, \beta) = P_{eist} + k_{ist}(1 - P_{eist})$. Dessa forma, Kappa é estimado resolvendo o terceiro conjunto de equações de estimação,

$$\mathbf{U}_3^*(\beta, \alpha^*) = \sum_{i=1}^{K} \left[\frac{\partial \mathbf{P}_{oi}}{\partial \alpha^*}\right]^t \mathbf{W}_i^{-1}(\mathbf{U_i} - \mathbf{P}_{oi}(\alpha^*, \beta)) = \mathbf{0} \tag{14}$$

em que $\mathbf{W}_i$ é a matriz de covariância de trabalho de $\mathbf{U}_i = (U_{i12}, U_{i13}, \ldots, U_{i34})$ tal que $\mathbf{W}_i = diag(\mathbf{P}_{oi}(1 - \mathbf{P}_{oi}))$, de modo que $\mathbf{U}_i - \mathbf{P}_{oi}$ representa os resíduos condicionais formados por todos os pares possíveis e distintos das associações entre safras.

Note que $P(Y_{is(k1)} = 1)$ se refere a média de todas as degustações das amostras oriundas da $s-$ésima safra classificadas na categoria de notas k_1, e que é função dos parâmetros de associações das degustações, bem como dos parâmetros do modelo marginal, α e β, respectivamente. Com isso, α^* será um vetor de estimativas dos parâmetros de associação para safras e P_{oist} é função de κ e P_{eist}. Para atender as restrições do espaço parâmetrico de κ usou-se a transformação de Fisher (KLAR; LIPSITZ; IBRAHIM, 2000):

$$k_{ist} = \frac{exp(\mathbf{z}'_{ist}\alpha^*) - 1}{exp(\mathbf{z}'_{ist}\alpha^*) + 1} \tag{15}$$

em que $\mathbf{z}'_{ist}$ é um vetor de covariáveis sensoriais para modelar κ e a concordância das respostas para a $i-$ésima degustação medida por κ e por iteração, computa-se $\hat{\alpha}^*$ como

$$\alpha^{*(m+1)} = \alpha^{*(m)} - \left(\sum_{i=1}^{K} \mathbf{C}_i{}^t \hat{\mathbf{W}}_i^{-1} \mathbf{C}_i\right)^{-1} \left(\sum_{i=1}^{K} \mathbf{U}_3^*(\alpha^{*(m)}, \beta(\alpha))\right),$$

em que $\mathbf{C}_i = \partial \mathbf{P}_{oi} / \partial \alpha^*$.

Convém ressaltar que não é cabível o ajuste de um modelo GEE usual, visto que as associações entre as degustações para o grupo de provadores não seriam contempladas e desta forma, desconsidera as variações existentes entre as notas das repetições das degustações para cada safra. Portanto, a proposta de modelagem torna-se efetiva no sentido de capturar outras variações que justifiquem as diferentes notas ao longo das safras. Nesse contexto, descreve-se uma estratégia para inserção das associações entre as múltiplas respostas ordinais repetidas.

2.7 Estratégia de modelagem para inserção das associações entre as múltiplas respostas ordinais repetidas

Segundo as especificações anteriores e mediante ao problema proposto, propôs-se que na construção do modelo marginal geral, a associação entre as degustações ocorridas em cada uma das safras sejam contemplada. Desse modo, as estimativas finais dos parâmetros desse modelo escrito em função das covariáveis altitudes, processamento, vertentes e genótipos, serão obtidas mediante a estratégia de modelagem que insere as associações entre as respostas originárias das desgustações realizadas por quatro provadores nos quatro genótipos ao modelo composto de 44 grupos distintos, conforme Tabela 1.

Desta forma, a estratégia proposta nesse trabalho é caracterizada na execução das seguintes etapas:

1. Ajusta-se modelos marginais para obtenção dos parâmetros de associação entre as respostas de cada provador ao longo das safras. Para cada safra, a estrutura de correlação comum para as desgustações será a permutável. Para tanto, tem-se quatro modelos marginais para obtenção dos vetores de parâmetros de associação entre degustações;

2. Obtidas as estimativas $\hat{\alpha} = (\hat{\alpha}_1^t, \hat{\alpha}_2^t, \hat{\alpha}_3^t, \hat{\alpha}_4^t)$, cujas dimensões variam de acordo com o número e tamanho dos grupos já citados anteriormente, constrói-se uma matriz de correlação de trabalho fixada e os parâmetros β do ajuste dos modelos marginais geral são obtidos;

3. após as etapas (1) e (2), obtem-se as quantidades matriciais, necessárias a serem utilizadas no processo iterativo para solução das equações de estimação(Equação 14), em que k_{ist} (equação 15), representa a medida de concordância das respostas entre as safras. Nessa etapa, as covariáveis envolvidas no processo iterativo para obtenção das estimativas $\hat{\alpha}^*$ serão as mesmas utilizadas para o ajuste do modelo geral.

Os parâmetros do modelo obtidos na etapa (2) são estimativas consistentes, quando a estrutura da matriz de correlação de trabalho para o modelo geral leva em consideração as associações entre as degustações realizadas por cada um dos provadores em cada uma das safras. E além disso, os dois conjuntos de resíduos usados nas etapas (1), (2) e (3) para cada parâmetro do modelo são estimados sob distintos conjuntos.

Os diferentes modelos apresentados neste artigo foram obtidos através do *software* R (R Core Team, 2015). Utilizando o pacote **geepack** para obtenção do ajuste dos dois primeiros conjuntos de equações de estimação (Halekoh, Højsgaard e Yan (2006), Yan e Fine (2004) e Yan (2002)).

3 RESULTADOS E DISCUSSÕES

3.1 Estudo descritivo das notas sensoriais dadas pelos provadores segundo safra e ajuste do modelo logito para categorias adjacentes

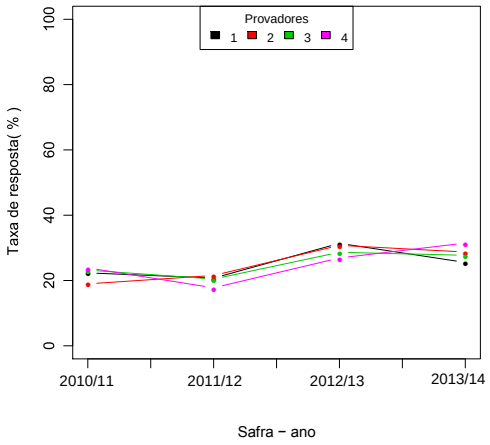
Os resultados ilustrados nas Figuras 2 e 3, evidenciam que as proporções de notas dadas pelos provadores foram mais homogêneas para a segunda categoria de notas(Figura 2(b)), enquanto que para a primeira categoria de notas finais, as menores proporções são dadas a terceira safra, indicando uma certa proximidade de respostas quanto a classificação dos cafés especiais obtidos na safra $2012/13$.

Em síntese, os cafés da terceira e quarta safras foram melhores classificados na segunda categoria de notas (Figura 3), sendo concordantes em um estudo similar utilizando uma modelagem probabilística de valores extremos proposto por Ferreira et al. (2016) no qual, considerou-se grupos de provadores não treinados em uma avaliação sensorial dos mesmos genótipos de cafés produzidos nessa mesma região.

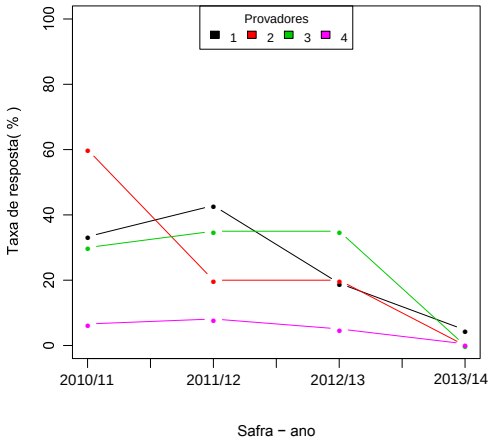
A primeira safra manteve-se com percentuais de notas sempre maiores que os da quarta safra na terceira categoria de notas, classe de notas elevadas.

Figura 2 – Perfis das notas finais por provadores para categorias de respostas em cada safra.

(a) Primeira categoria

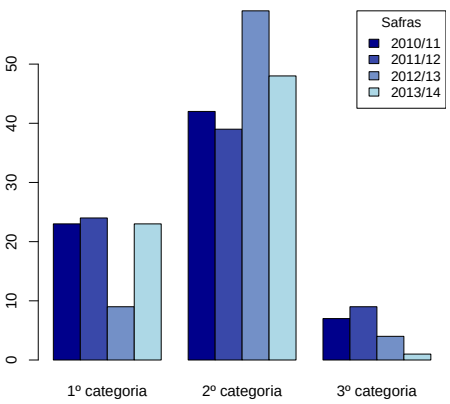


(b) Segunda Categoria

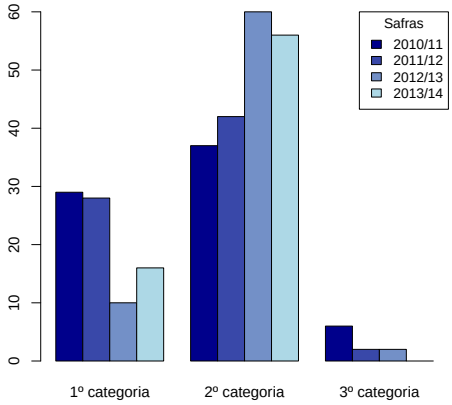


(c) Terceira categoria

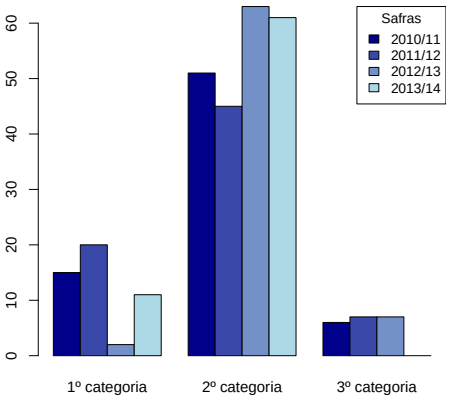
Figura 3 – Gráfico em barras para categorias de notas de cada um dos provadores.



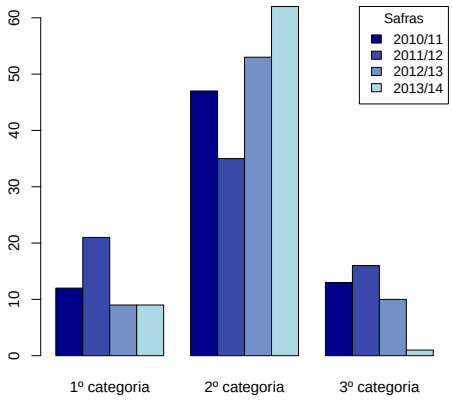
(a) Primeiro provador



(b) Segundo provador



(c) Terceiro provador



(d) Quarto provador

Dada as respostas categóricas, cujas categorias possuem uma ordenação natural e tendo interesse em identificar qual será a chance da classificação da nota final dada pelos provadores estar em uma determinada categoria, considerou-se o modelo proposto na equação (3) de categorias adjacentes de modo que, $logito_1 : log(\pi_1/\pi_2) = \beta_{01} + \beta_2 \text{provador} + \beta_3 \text{safra} + \beta_4 \text{genótipo}$ representa o log da chance de classificação da nota final dada aos cafés especiais estar na categoria de notas baixas(primeira classe) em comparação com as demais categorias; já o $logito_2 : log(\pi_2/\pi_3) = \beta_{02} + \beta_2 \text{provador} + \beta_3 \text{safra} + \beta_4 \text{genótipo}$ representa o log da chance da classificação das notas estarem na primeira ou segunda classes em comparação à categoria de notas mais elevadas.

Após o ajuste do modelo, considerando cada uma das variáveis envolvidas e com o teste da razão de verossimilhança, identificou-se que a safra não atente ao pressuposto de proporcionalidade, ou seja, a estimativa do coeficiente da variável safra(β_3) não é a mesma para os logitos 1 e 2. E assim, o log das chances não é idêntico entre as categorias, ou seja, o modelo assume que existem observações que possuem variância heterocedásticas e que a variável safra oscilará de acordo com a categoria de resposta. Dessa forma, utilizou-se o modelo logitos proporcionais parciais, cujas estimativas encontram-se na Tabela 5.

Tabela 5 – Estimativas dos parâmetros para o modelo de chances proporcionais parciais

Coeficientes	Estimativas	Erro padrão	p-valor
β_{01}	-0,1309	0,1857	0,48091
β_{02}	2,9038	0,2613	2×10^{-16}
Safra($logito_1$)			
2011/12	0,4659	0,1994	0,01944
2012/13	-1,3908	0,2485	$2,18 \times 10^{-08}$
2013/14	-0,5946	0,2125	0,00513
Safra($logito_2$)			
2011/12	-0,1939	0,2840	0,49493
2012/13	0,4702	0,3029	0,12055
2013/14	2,9562	0,7405	$6,54 \times 10^{-05}$
Provador			
2	0,4348	0,1689	0,01006
3	-0,4274	0,1699	0,01188
4	-0,6716	0,1718	$9,27 \times 10^{-05}$
Genótipo			
Catuaí amarelo	-1,2339	0,1344	2×10^{-16}
Acaiá	-1,8771	0,4234	$9,29 \times 10^{-06}$
Mundo novo	-0,4054	0,3926	0,30185

De acordo com os resultados da Tabela 5, em que os efeitos do genótipo Bourbon amarelo e safra 2010/11 são confundidos com o intercepto, observou-se que as amostras degustadas do genótipo Bourbon amarelo oriundas da safra 2013/14, tem maiores chances de serem classificadas em categorias de notas maiores em relação às amostras provenientes da safra 2010/11. Vale ressaltar que na safra 2010/11, há uma maior variedade de genótipos que foram avaliados, ao passo que na safra 2013/14 foram somente dois genótipos degustados. Contudo, a razão de chances entre as amostras de Catuaí amarelo e Bourbon amarelo pode ser estimadas em $e^{-1,2339} = 0,2911$. Assim, as amostras de Catuaí amarelo provenientes da safra 2010/11 tem menores chances de serem classificadas nas categorias de maiores notas.

3.2 Os modelos marginais

Dado o propósito de obter estimativas mais sensíveis ao possível efeito de safra, realizou-se o ajuste do modelo de acordo com a equação (5) com interceptos não constantes em relação às safras. Assim, o modelo cujas notas estejam nas categorias acima da classe de notas baixas, será:

$$\begin{aligned} logit(E(Y_{ij_{h(1)}})) = {} & 0,2395(\text{provador}_1) + 0,2518(\text{provador}_2) + 0,42020(\text{provador}_3) \\ & -0,2567(\text{provador}_4) + 0,7889(\text{genótipo}_1) + 0,7293(\text{genótipo}_2) \\ & +0,7758(\text{genótipo}_3) + 0,87014(\text{genótipo}_4) - 0,03704(\text{safra}_1) \\ & -0,64983(\text{safra}_2) - 0,3637(\text{safra}_3) - 0,6572(\text{safra}_4) \end{aligned}$$

e para as notas acima da segunda categoria de respostas:

$$\begin{aligned} logit(E(Y_{ij_{h(2)}})) = {} & 0,2395(\text{provador}_1) + 0,2518(\text{provador}_2) + 0,42020(\text{provador}_3) \\ & -0,2567(\text{provador}_4) + 0,7889(\text{genótipo}_1) + 0,7293(\text{genótipo}_2) \\ & +0,7758(\text{genótipo}_3) + 0,87014(\text{genótipo}_4) - 4,1057(\text{safra}_1) \\ & -4,4801(\text{safra}_2) - 4,6359(\text{safra}_3) - 4,8265(\text{safra}_4) \end{aligned}$$

Fixada a $j-$ésima safra e assumindo a estrutura de correlação permutável, foram obtidas conforme equação 9, as correlações entre categorias de notas para cada uma das safras (Tabela 6). Com relação a Correlação($\rho_j(1,2)$), observou-se que a variação global das respostas categóricas foi melhor explicada quando se realiza o agrupamento por safras. Assim, pode-se afirmar que a classificação das notas acima das primeira e segunda categorias é uma característica das safras ano $2010/11$ e $2012/13$.

Resultado semelhante é observado quando se refere às probabilidades das safras serem classificadas nas categorias de maiores notas. Na Table 6 encontram-se as razões de chances estimadas pela equação (8) e respectivas probabilidades.

Segundo a Tabela 6, na ocasião em que as amostras são provenientes das safras $2010/11$ e $2012/13$ as probabilidades de haverem notas acima da primeira categoria são maiores para essas duas safras, bem como as chances se mostram ser maiores para as duas situações descritas na Tabela 6, indicando que as notas para essas safras tem duas vezes mais chances de serem classificadas em classes de notas mais elevadas.

Tabela 6 – Probabilidades estimadas do modelo com intercepto não constante de categorias de notas, razão de chances e correlação para todas as safras, segundo grupo de provadores e genótipos.

Situação em que as notas estão nas categoria		Safras			
		2010/11	2011/12	2012/13	2013/14
Acima de 82 pontos	Probabilidades	0,7293	0,5821	0,6968	0,4890
	Chances	2,6953	1,3929	2,2987	0,9570
Acima de 91 pontos	Probabilidades	0,0441	0,0293	0,0311	0,0146
	Chances	0,0460	0,0302	0,0321	0,0148
	Correlação($\rho_j(1,2)$)	0,4587	0,1642	0,3937	-0,0219

Estas observações confirmam as análises gráficas das Figuras 2 e 3 e, além disso, reforçam a identificação de que a safra 2012/13 foi melhor classificada na segunda categoria de notas e que a safra 2013/14, teve o menor desempenho na probabilidade e para as situações da Tabela 6.

Seguindo a estratégia sugerida nesse trabalho e considerando as covariáveis altitude (inferiores a 1.000m, entre 1.000 e 1.200m e acima de 1.200m), dois tipos de processamento (Natural ou via seca e cereja descascada ou via úmida), vertente(Sol e sombra) e genótipos(Bourbon amarelo, Catuaí amarelo, Acaiá vermelho e Mundo novo), ajustou-se modelos marginais para cada safra assumindo a matriz de correlação de trabalho permutável e na primeira etapa da estratégia de modelagem as associações obtidas foram: $\hat{\alpha} = (0,09559, 0,01208, 0,01440, 0,02698)$.

De posse das estimativas $\hat{\alpha}$, aplicou-se a segunda etapa da estratégia de modelagem: ajuste do modelo marginal geral. As estimativas dos parâmetros constam na Tabela 7.

Tabela 7 – Estimativas dos parâmetros dos quatro modelos marginais, obtidos separadamente, para o conjunto de todos os provadores nas quatro safras.

Coeficientes	Estimativas por safras dos modelos marginais individuais				Modelo geral
	2010/11	2011/12	2012/13	2013/14	
Intercepto	-1,4983	-1,1744	-0,17134	-0,8839	-0,8719
Altitude					
$1.000 - 1.200$	0,5481	0,0982	-0,0482	0,1362	0,1749
> 1.200	1,8554	0,5147	0,1493	0,4620	0,7022
Vertente					
Sombra	-0,0772	0,0465	-0,2452	0,0444	-0,0476
Processamento					
Cereja descascada Via úmida	-0,1873	-0,1411	-0,1272	-0,0148	-0,1137
Genótipo					
Catuaí amarelo	0,7765	1,1022	0,5448	0,5157	0,6617
Acaiá	0,9943	2,2264	NA	NA	1,0976
Mundo novo	0,6964	NA	NA	NA	0,3349

NA: refere-se ao genótipo não avaliado na safra.

Nesta ocasião, verificou-se que a razão de chances entre as altitudes superiores e inferiores a $1.200m$ é estimada em $e^{1,8554}$. Ou seja, as amostras de cafés especiais do genótipo Bourbon amarelo da safra $2010/11$, provenientes de altitudes superiores a $1.200m$ possuem aproximadamente 6 vezes a chance de serem classificadas com notas maiores que as amostras oriundas de altitudes inferiores a $1.200m$. Notou-se que para os genótipos avaliados da safra $2011/12$, Catuaí amarelo e Acaiá, apresentam maiores chances de serem classificados com maiores notas que o genótipo Bourbon amarelo. Porém, na terceira e quarta safras, na ocasião em que há somente dois genótipos degustados, as amostras de cafés especiais de Bouborn Amarelo ourindas de altitudes superiores a 1.200m, cujo processamento é o natural apresentam maiores chances de serem classificadas nas categorias de maiores notas.

No que segue a proposta desse trabalho e na tentativa de identificar mudanças das notas dadas ao genótipos ao longo das safras, segundo altitude, processamento e vertentes, utilizaremos as equações (14) e (15), executando assim, o terceiro passo da estratégia proposta nesse artigo.

Conforme citado anteriormente, as estimativas $\hat{k}_{ist}$ medem o grau de concordância entre as safras e serão dadas segundo o grupo de covariáveis já citadas. Vale ressaltar que há 288 avaliações(degustações) por safras, dessa forma a Tabela 8 fornece os valores médios de $k_{ist}, i = 1,2,\ldots,288$, $s,t = 1,2,3,4$, $s < t$ correspondentes a concordância das notas dadas aos cafés especiais.

Tabela 8 – Estimativas médias dos valores Kappa das medidas de concordância das degustações para todas as combinações entre safras.

	Associações safra - valores de $\bar{\kappa}$					
Covariáveis	1 - 2	1 - 3	1 - 4	2 - 3	2 - 4	3 - 4
Altitude						
$1.000 - 1.200$	0,3331	-0,3227	-0,2803	0,0033	0,3333	-0,3327
> 1.200	0,3331	-0,3227	-0,2803	0,0033	0,3333	-0,3327
Vertente						
Sombra	0,4997	-0,4841	-0,4205	0,0049	0,5000	-0,4990
Processamento						
Cereja descascada Via úmida	0,4997	-0,4841	-0,4205	0,0049	0,5000	-0,4990
Genótipo						
Catuaí amarelo	0,4441	-0,4303	-0,3738	0,0044	0,4444	-0,4436
Acaiá	0,0555	-0,0537	-0,0467	0,0005	0,0555	-0,0554

Os resultados apontam que as amostras identificadas pela vertente sombra, ourindas das safras $2010/11-2011/12$ e $2011/12-2013/14$, apresentam um grau de concordância moderados. O mesmo ocorre para as amostras de cafés especiais Catuaí amarelo e as amostras identificadas pelo processamento via úmido. O destaque para valores de $\bar{\kappa}$ próximos de zero, foram os referentes as safras $2011/12-2012/13$ indicando que a concordância entre as amostras de cafés especiais provenientes das safras $2011/12$ e $2012/13$ são esperadas pelo acaso, ou seja, não há indícios de que as notas dadas as amostras das referidas safras segundo o conjunto de covariáveis altitudes, vertentes, processamento e genótipos apresentam um indicativo de que as safras possuem características comuns.

Vale ressaltar que os valores negativos de $\bar{\kappa}$ indicam discordância, porém nada se pode dizer com relação ao grau de discordância.

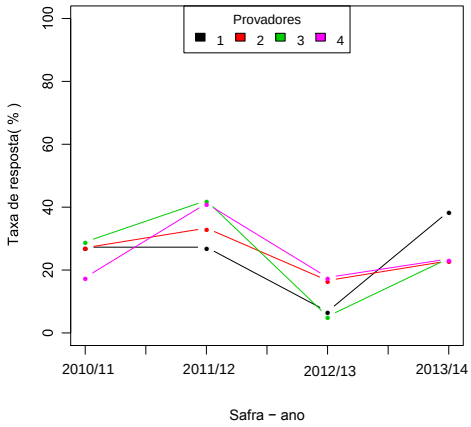
3.3 Estudo descritivo das notas sensoriais dadas pelos provadores segundo safra e ajuste do modelo logito para categorias adjacentes - Análise para cinco categorias de notas

Para as covariáveis altitude e processamento, realizou-se análise do comportamento das respostas ao longo das quatro safras, sob a perspectiva de cinco classes e dois genótipos, a saber: Bourbon Amarelo e Catuaí amarelo. As respostas foram categorizadas da seguinte forma: (1 : notas finais < 80; 2 : entre $80-82$ inclusive; 3 : $82-86$ inclusive; 4 : notas finais entre $86-89$ inclusive e, 5 : notas finais > 89), avaliadas para o $i-$ésimo provador na $j-$ésima safra, $i, j = 1,2,3,4$.

Os resultados ilustrados nas Figuras 4, 5 e Figura 6, evidenciam que as proporções de notas dadas pelos quatro provadores aos genótipos Bourbon e Catuaí amarelos, foram mais homogêneas para notas entre 80 e 82 (Figura 4(b)) entre todas as safras, enquanto que para a quarta categoria de notas(Figura 4(d)) o destaque é para safra $2012/13$. Notou-se que a quarta safra foi melhor classificada com notas finais inferiores a 89, ao passo que a safra $2011/12$ com notas inferiores a 86.

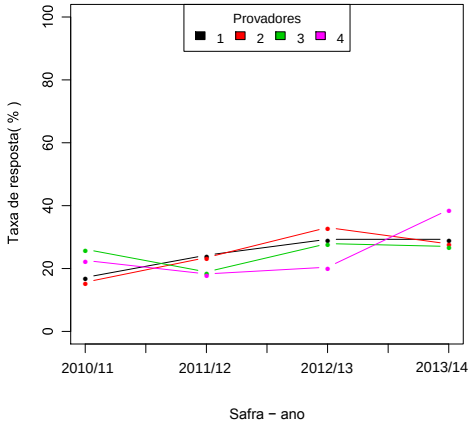
Em síntese, os cafés da terceira e quarta safras foram melhores classificados na terceira categoria de notas (Figura 6), resultados concordantes quando se realiza análise para três categorias de notas.

Figura 4 – Perfis das notas finais por provadores para cinco categorias de respostas em cada safra.

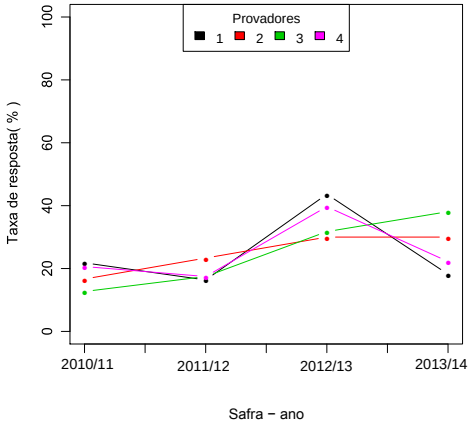


(a) Primeira categoria

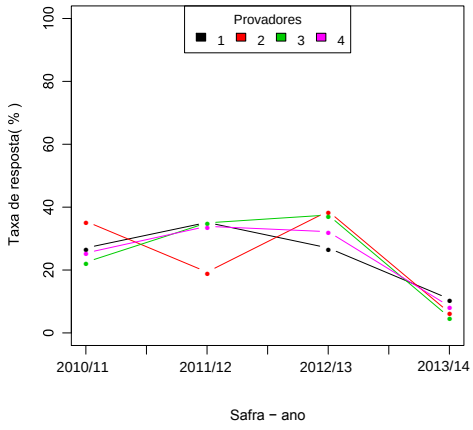
(b) Segunda categoria



(c) Terceira categoria

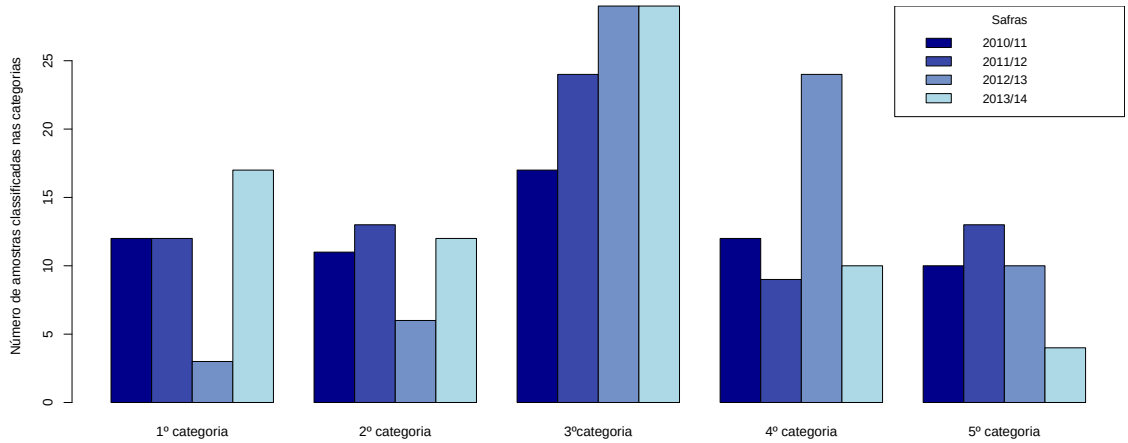


(d) Quarta categoria

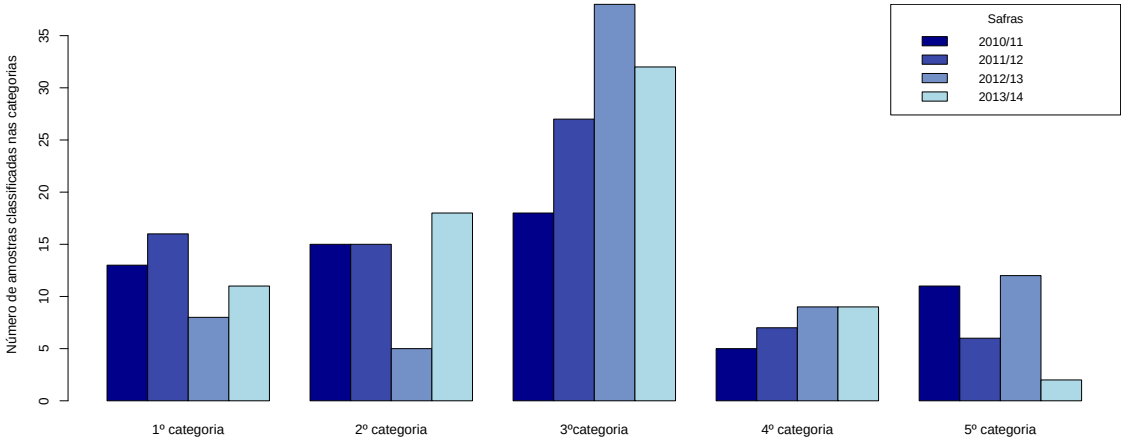


(e) Quinta categoria

Figura 5 – Gráfico em barras para categorias de notas do primeiro e segundo provadores.

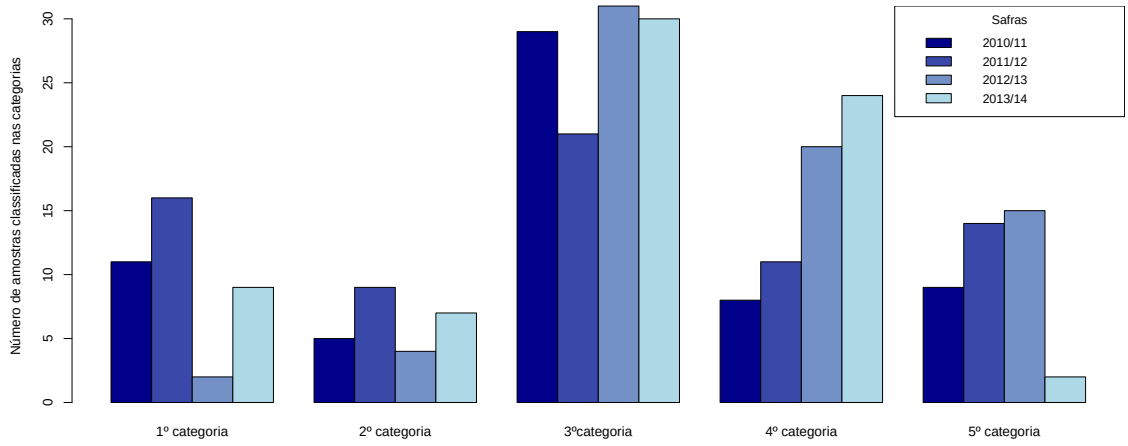


(a) Primeiro provador

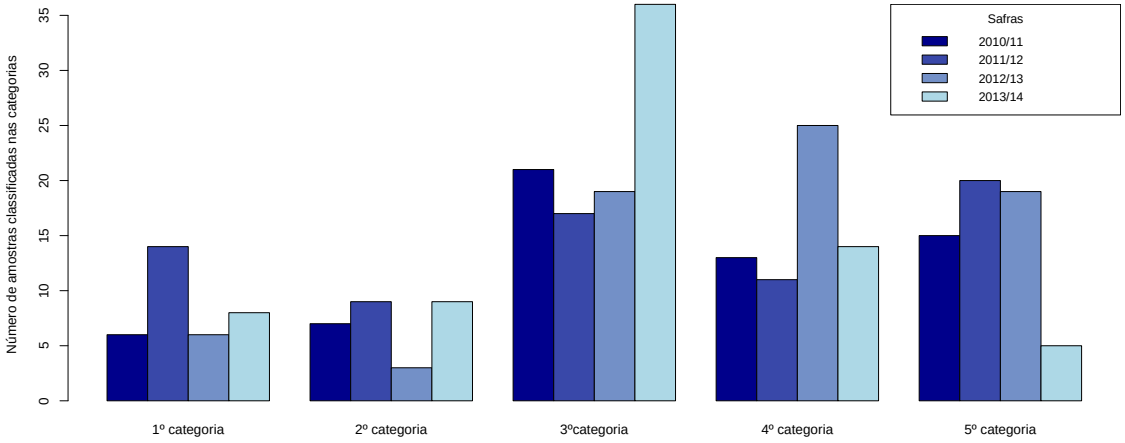


(b) Segundo provador

Figura 6 – Gráfico em barras para categorias de notas do terceiro e quarto provadores.



(a) Terceiro provador



(b) Quarto provador

Dadas as respostas categóricas, cujas categorias possuem uma ordenação natural e tendo interesse em identificar qual será a chance da classificação da nota final dada pelos provadores estar em uma determinada categoria, sob avaliação das covariáveis provador, safra, altitude, processamento e genótipo, considerou-se o modelo logit de categorias adjacentes:

$$log\frac{\pi_k(\mathbf{x})}{\pi_{k+1}(\mathbf{x})} = \theta_k + \beta_1 \text{provador}_{ik} + \beta_2 \text{safra}_{ik} + \beta_3 \text{altitude}_{ik} + \beta_4 \text{processamento}_{ik} + \beta_5 \text{genótipo}_{ik}, \quad k = 1,2,3,4. \tag{16}$$

com efeitos β comuns para cada um dos 4 modelos logit.

Após o ajuste, com o teste da razão de verossimilhança, identificou-se que somente as covariáveis provador e genótipo atendem ao pressuposto de proporcionalidade, ou seja, a contribuição para a respostas em cada um dos logitos é a mesma.

Para as covariáveis que não atendem ao pressuposto de proporcionalidade, entende-se que o log das chances não é idêntico entre as categorias, assumindo que no modelo existem observações que possuem variância heterocedásticas e que as variáveis não proporcionais entre os logitos, oscilaram de acordo com a categoria de resposta.

Dessa forma, utilizou-se o modelo logitos proporcionais parciais, cujas estimativas são apresentadas na Tabela 9

Tabela 9 – Estimativas dos parâmetros para o modelo de chances proporcionais parciais para cinco categorias de notas

Coeficientes	$Log(\pi_k/\pi_{k+1})$ (Razão de chances das estimativas)			
	$log(\pi_1/\pi_2)$	$log(\pi_2/\pi_3)$	$log(\pi_3/\pi_4)$	$log(\pi_4/\pi_5)$
Intercepto	$0,8682^*$ (2,3828)	$0,7620^*$ (2,1426)	$2,8046^*$ (16,5206)	$1,6938^*$ (5,4403)
Provador				
2	$0,1827^*$ (1,2004)	$0,1827^*$ (1,2004)	$0,1827^*$ (1,2004)	$0,1827^*$ (1,2004)
3	$-0,1428$ (0,8669)	$-0,1428$ (0,8669)	$-0,1428$ (0,8669)	$-0,1428$ (0,8669)
4	$-0,2898^*$ (0,74839)	$-0,2898^*$ (0,7483)	$-0,2898^*$ (0,7483)	$-0,2898^*$ (0,7483)
Safra				
2011/12	$0,1435$ (1,15438)	0.1112 (1,11768)	$-0,0420$ (0,9588)	$-0,0338$ (0,9666)
2012/13	$-0,1111$ (0,8947)	$-1,2091^*$ (0,2984)	$-0,6186^*$ (0,5386)	$0,5346^*$ (1,7068)
2013/14	$-0,1033$ (0,9018)	$-0,2158$ (0,8058)	$-0,0431$ (0,9577)	$1,9256^*$ (6,8598)
Altitude				
$1.000 - 1.200$	$-0,1872$ (0,8292)	$-0,3286$ (0,7198)	$-0,4720^*$ (0,6237)	$-0,1980$ (0,8203)
> 1.200	$-0,0152$ (0,9849)	$-0,7471$ (0,47372)	$-0,8976$ (0,4075)	$-1,4722$ (0,2293)
Processamento				
Cereja descascada	$0,5984^*$	-0,1359	$-0,2165$	$0,5742^*$
Via úmida	(1,8192)	(0,8729)	(0,8053)	(1,7758)
Genótipo				
Catuaí amarelo	$-0,7820^*$ (0,4574)	$-0,7820^*$ (0,4574)	$-0,7820^*$ (0,4574)	$-0,7820^*$ (0,4574)

De acordo com a Tabela 9, a covariável processamento melhora significativamente as chances das notas dadas aos cafés especiais, estarem na primeira e quarta categorias de notas. Observou-se que a covariável safra, contribuiu positivamente para que as chances dos cafés especiais serem classificados na quarta categoria, sejam maiores que as chances de estarem na quinta categoria de notas. De maneira geral, a razão das chances de classificação das notas são maiores para a terceira e quarta categorias de notas.

Em comparação com a análise realizada com três categorias de notas, a maior chance ficou para a segunda categoria, ou seja, os cafés especiais foram melhor classificados com as notas finais entre 82 e 91. Convém ressaltar que para a análise de cinco categorias, apesar da redução do número de genótipo avaliados, também indicou que os cafés especiais Bourbon Amarelo e Catuaí amarelo foram melhor classificados com notas finais entre 82 e 86 e, 86 e 89 para as terceira e quarta categorias respectivamente. Tal comparação deve realizada com cautela, visto que para a análise de três categorias, no modelo logit de categorias adjacentes não levou-se em consideração as covariáveis altitude e processamento.

De acordo com a equação (5), ajustou-se um modelo com interceptos não constantes em relação às safras e calculou-se as probabilidades estimadas da classificação das notas acima das categorias $k = 1,2,3,4$.

Na ocasição em que as amostras são provenientes da safra $2010/11$ a probabilidade de haverem notas acima da quarta categoria de notas(notas finais entre 86 e 89) são maiores, bem como a safra $2013/14$ foi melhor classificada na terceira categoria de notas, ao passo que de maneira geral, os genótipos avaliados apresentaram probabilidades maiores de classificação para segunda categoria de notas(Tabela 10).

Tabela 10 – Probabilidades estimadas do modelo com intercepto não constante para cinco categorias de notas segundo grupo de provadores e genótipos

Situação em que as notas estão acima da categoria	Safras			
	2010/11	2011/12	2012/13	2013/14
< 80	0,8116	0,7583	0,8047	0,8707
$80 - 82$	0,6378	0,5262	0,6852	0,7551
$82 - 86$	0,2527	0,1284	0,2361	0,3629
$86 - 89$	0,6825	0,0592	0,0755	0,1536

4 CONCLUSÃO

As associações entre as degustações para a interação grupo de provadores e genótipos avaliados, considerando as notas das avaliações fornecidas ao longo das safras e as covariáveis altitude, vertente e processamento, foram explicadas pelo modelo geral. Concluiu-se que a estratégia de modelagem foi adequada por discriminar as diferenças entre as categorias de notas mais elevadas e de menores notas. Observou-se que os grupos de provadores possuem similaridades de notas para a safra, ano $2012/13$, e que segundo as covariáveis, as safras $2010/11 - 2012/13$, $2010/11 - 2013/14$ e $2012/13 - 2013/14$ discordam entre si. Contudo, as estimativas médias das medidas de concordância entre as safras $2010/11 - 2011/12$ e $2011/12 - 2013/14$ indicam concordância moderada segundo o conjunto de covariáveis vertente do tipo sombra, processamento cereja descascada e genótipo Catuaí amarelo.

Ressalta-se que a aplicabilidade dessa nova estratégia poderá ser ineficaz caso o experimento apresente grupos de dimensões maiores, pois, computacionalmente, o uso das equações de estimação generalizadas para dados ordinais é limitada.

Agradecimentos

Os autores agradecem à FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e INCT (Instituto Brasileiro de Ciência e Tecnologia do Café) pelo suporte financeiro.

5 REFERÊNCIAS

AGRESTI, A. **Categorical Data Analysis (Chapter 8, section 8.3.4)**. [S.l.]: Wiley John + Sons, 2013. ISBN 0470463635.

BOREM, F. M. **Projeto protocolo de identidade, qualidade e rastreabilidade para embasamento da indicacão geográfica dos cafés da mantiqueira**. [S.l.], 2007.

CAREY, V.; ZEGER, S. L.; DIGGLE, P. Modelling multivariate binary data with alternating logistic regressions. **Biometrika**, Biometrika Trust, v. 80, n. 3, p. 517–526, 1993.

CLAYTON, D. Repeated ordinal measurements: A generalised estimating equation approach. **Medical Research Council Biostatistics Unit Technical Report. Cambridge, England**, 1992.

FERREIRA, H. A. et al. Selecting a probabilistic model applied to the sensory analysis of specialty coffees performed with consumer. **IEEE Latin America Transactions**, v. 14, n. 3, p. 1507–1512, mar. 2016. ISSN 1548-0992.

FITZMAURICE, G. M.; LAIRD, N. M. A likelihood-based method for analysing longitudinal binary responses. **Biometrika**, Biometrika Trust, v. 80, n. 1, p. 141–151, 1993.

GANGE, S. et al. Analysis of correlated ordinal measures with ophthalmic applications. **University of Wisconsin, Technical report**, 1993.

GONIN, R. et al. Regression modelling of weighted κ by using generalized estimating equations. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 49, n. 1, p. 1–18, 2000.

HALEKOH, U.; HØJSGAARD, S.; YAN, J. The r package geepack for generalized estimating equations. **Journal of Statistical Software**, v. 15, n. 2, p. 1–11, 2006.

HEAGERTY, P. J.; ZEGER, S. L. Marginal regression models for clustered ordinal measurements. **Journal of the American Statistical Association**, Taylor & Francis, v. 91, n. 435, p. 1024–1036, 1996.

KLAR, N.; LIPSITZ, S. R.; IBRAHIM, J. G. An estimating equations approach for modelling kappa. **Biometrical Journal**, Wiley Online Library, v. 42, n. 1, p. 45–58, 2000.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. **Biometrika**, Biometrika Trust, v. 73, n. 1, p. 13–22, 1986.

LINGLE, T. R. **The coffee cupper's handbook: a systematic guide to the sensory evaluation of coffee's flavor**. [S.l.]: Specialty Coffee Association of America Long Beach, CA, 2011.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society A**, v. 135, p. 370–84, 1972.

PRENTICE, R. L.; ZHAO, L. P. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. **Biometrics**, JSTOR, p. 825–839, 1991.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2015. Disponível em: <https://www.R-project.org/>.

WILLIAMSON, J.; KIM, K. A global odds ratio regression model for bivariate ordered categorical data from ophthalmologic studies. **Statistics in medicine**, v. 15, n. 14, p. 1507–1518, 1996.

WILLIAMSON, J. M.; KIM, K.; LIPSITZ, S. R. Analyzing bivariate ordinal data using a global odds ratio. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 90, n. 432, p. 1432–1437, 1995.

WILLIAMSON, J. M.; MANATUNGA, A. K.; LIPSITZ, S. R. Modeling kappa for measuring dependent categorical agreement data. **Biostatistics**, v. 1, n. 2, p. 191–202, 2000.

YAN, J. Geepack: yet another package for generalized estimating equations. **R news**, v. 2, n. 3, p. 12–14, 2002.

YAN, J.; FINE, J. Estimating equations for association structures. **Statistics in medicine**, Wiley Online Library, v. 23, n. 6, p. 859–874, 2004.

ZHAO, L. P.; PRENTICE, R. L. Correlated binary regression using a quadratic exponential model. **Biometrika**, Biometrika Trust, v. 77, n. 3, p. 642–648, 1990.

ARTIGO 2

Critério de seleção da matriz de trabalho em função das estimativas limitantes da matriz de covariância de dados correlacionados em GEE

Artigo redigido conforme normas da Universidade Federal de Lavras
(versão preliminar)

ARTIGO 2

Critério de seleção da matriz de trabalho em função das estimativas limitantes da matriz de covariância de dados correlacionados em GEE

RESUMO

A modelagem de equações de estimação generalizadas(GEE), utilizada na análise de dados longitudinais seja em variáveis contínuas ou discretas, requer necessariamente a especificação, à priori, de uma matriz de correlação em seu processo iterativo, para obtenção das estimativas dos parâmetros de regressão. Tal matriz é denominada como matriz de correlação de trabalho, e a sua incorreta especificação, pode produzir estimativas menos eficientes para os parâmetros do modelo. Decorrente a esse fato, este trabalho tem por objetivo propor um critério de seleção da matriz de correlação de trabalho, baseado nas estimativas da matriz de covariância de respostas correlacionadas provenientes dos valores limitantes das estimativas dos parâmetros de associação. Para validação do critério, utilizou-se estudos via simulação considerando respostas correlacionadas normais e binárias. Em comparação a alguns critérios existentes na literatura, concluiu-se que o critério proposto resultou em um melhor desempenho, quando a estrutura de correlação para matriz de correlação de trabalho permutável foi considerada como estrutura verdadeira nas amostras simuladas, e para grandes amostras, o critério proposto apresentou comportamento similar ao demais critérios, resultando em maiores taxas de acerto.

Palavras-chave: Critério. Equação de estimação generalizada. Matriz de correlação de trabalho. Dados correlacionados.

ARTICLE 2

Criterion of the selection of a working correlation structure in function of limiting estimates of the covariance matrix for correlated data in the GEE

ABSTRACT

The modeling of generalized estimation equations used in the analysis of longitudinal data whether in continuous or discrete variables, necessarily requires the prior specification of a correlation matrix in its iterative process to obtain the estimates of the regression parameters. Such an array is called a working correlation matrix and its incorrect specification produces less efficient estimates for the model parameters. Due to this fact, this work aims to propose a criterion of selection of the work correlation matrix, based on the estimates of the covariance matrix of correlated responses coming from the limiting values of the association parameter estimates. For validation of the criterion, we used simulation studies considering normal and binary correlated responses. Compared to some criteria in the literature, it was concluded that the proposed criterion resulted in a better performance when the correlation structure for exchangeable working correlation matrix was considered as true structure in the simulated samples for large samples, the proposed criterion presented similar behavior to the other criteria, resulting in higher hit rates.

Keywords: Criterion. Generalized estimation equation. Working correlation structure. Correlated data

1 INTRODUÇÃO

Estudos longitudinais são caracterizados por permitirem incorporar no modelo, o desenvolvimento individual de uma característica de interesse ao longo do tempo, em conexão com um conjunto de covariáveis. Nesse contexto, múltiplas medidas sob um mesmo indivíduo ao longo do tempo ou em diversas ocasiões de observações produzem respostas correlacionadas, e para tanto, há necessidade de descrever a associação existente entre tais respostas. Modelos marginais são utilizados para modelar a resposta média, marginalizada, em cada tempo ou ocasição, considerando as covariáveis de efeito fixo e a incorporação da associação entre as respostas longitudinais. Assim, para esses modelos, assumindo independência entre os indivíduos, fornecem estimativas do efeito das covariáveis na esperança marginal da variável resposta.

Dentre as metodologias da literatura, a abordagem GEE(equações de estimação generalizadas), introduzida por Liang e Zeger (1986), fundamentada em quase-verossimilhança, não pressupõe a especificação completa da distribuição multivariada das respostas repetidas, porém requer a identificação dos dois primeiros momentos. O método depende fortemente do uso da matriz de correlação de trabalho que a princípio é escolhida de forma arbitrária pelo pesquisador.

A descrição dos dados inicia-se por considerar i indivíduos, $i = 1,2,\ldots,K$, $\mathbf{Y}_i = \mathbf{y}_{it}$ um vetor de n_i medidas repetidas tomadas em t ocasiões, $1 \leq t \leq n_i$. As respostas podem ser variáveis contínuas ou discretas em que se assume uma combinação linear do vetor de covariáveis $\mathbf{X}_{it}$, $p \times 1$. Sejam $\mathbf{C}_i(\rho)$ e $\Sigma_i(\rho)$ as respectivas matrizes $n_i \times n_i$ de correlações e covariâncias verdadeiras de $\mathbf{Y}_i$ $(i = 1,\ldots,K)$, que usualmente são desconhecidas. O parâmetro de correlação, ρ, caracteriza completamente $\mathbf{C}_i(\rho)$.

Para o caso em que $\mathbf{C}_i(\rho)$ é desconhecida, Liang e Zeger (1986) propuseram uma metodologia com base em quase-verossimilhança para obtenção das estimativas dos parâmetros β, baseada na matriz de correlação de trabalho $\mathbf{R}_i(\alpha)$, em que α é um parâmetro de correlação, que na prática, também é desconhecido. Desta forma, considerando $\hat{\alpha}$ um estimador consistente de α, cujas estimativas são obtidas pelo método dos momentos, tais que o vetor de médias $\mu_{it} = g(\mathbf{X}_{it}^T\beta)$ e variância $\phi\sigma_{it}^2$, no qual ϕ refere-se ao parâmetro escalar desconhecido que geralmente é fixado em $\phi = 1$ e $\mathbf{A}_i = diag(\sigma_{it}^2)$. A matriz de covariância de trabalho de $\mathbf{Y}_i$ é definida por $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}$.

Com essas especificações, as estimativas de quase-verossimilhança para β, são obtidas pela solução de:

$$\mathbf{U}(\beta,\alpha)=\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}(\mathbf{Y}_i-\mu_i)=\mathbf{0}, \tag{2}$$

em que $\mathbf{D}_i=\partial\mu_i/\partial\beta^T$. Dessa forma, a matriz de covariância estimada de $\mathbf{Y}_i$, $\hat{\mathbf{V}}_i$ será dada em função da matriz de correlação estimada, $\mathbf{R}_i(\hat{\alpha};\hat{\beta})$.

Dado que $\hat{\alpha}(\beta)$ é um estimador consistente para α, Zhao, Prentice e Self (1992) em estudos de simulação, sob a suposição do verdadeiro valor para estimativa dos parâmetros β, avaliaram a eficiência dos estimadores $\hat{\beta}_G$(baseado na estrutura permutável ou AR(1)) e $\hat{\beta}_I$(baseado na estrutura identidade) para verificação da especificação incorreta da matriz de covariância com base na matriz de correlação de trabalho estimada, $\mathbf{R}(\hat{\alpha};\hat{\beta}_G)$.

Sutradhar e Das (2000) considerando que a eficiência computacional se baseia na matriz $\mathbf{R}(\alpha_0(\rho))$, em que $\alpha_0(\rho)$ é o valor limitante das estimativas de α, realizaram uma avaliação computacional com propósito de comparar a eficiência de $\hat{\beta}_I$ e posteriormente efetuaram análise comparativa entre $\hat{\beta}_G$ e $\hat{\beta}_T$(verdadeiro estimador de quase-verossimilhança de β), sob $\mathbf{R}(\alpha_0(\rho))$ e confirmaram resultados apresentados em Sutradhar e Das (1999) de que, os estimadores obtidos sob a suposição de independência produziram estimativas menos eficientes quando comparado com $\hat{\beta}_G$, e além disso que a eficiência das estimativas dos parâmetros β, depende da especificação da verdadeira estrutura de correlação de trabalho e da magnitude dos parâmetros de correlação ρ para obtenção das estimativas de α_0.

Decorrente a importância de especificar a estrutura da matriz de correlação de trabalho corretamente, de modo a garantir uma melhor eficiência nas estimativas dos parâmetros β, na literatura, encontram-se propostos vários critérios de seleção para melhor escolha da referida matriz. Rotnitzky e Jewell (1990) em análises da extensão do teste qui-quadrado para testar hipóteses sob um conjunto de parâmetros, examinaram o comportamento assintótico da estatística de Wald sob o pressuposto da especificação da verdadeira matriz de correlação de trabalho, aplicada em estudos de associações dentro dos grupos. Na literatura é dito "Rotnizky and Jewell's criterion (RJC)".

Posteriormente, Hin, Carey e Wang (2007) descreveram o critério RJC para a seleção da estrutura de correlação de trabalho. Pan (2001) propôs uma abordagem sob a modificação do AIC para seleção de modelos na abordagem GEE, e consequentemente como critério de seleção para estrutura de correlação de trabalho chamando de "quasi-likelihood under the independence model criterion" (QIC).

Hin e Wang (2009) propuseram usar metade do segundo termo de QIC para selecionar a estrutura de correlação de trabalho em GEE, originando o Critério de Informação de Correlação (CIC). Gosho, Hamada e Yoshimura (2011) propuseram uma medida da discrepância entre o estimador da matriz de covariância e uma matriz de covariância especificada que considera a soma dos elementos da diagonal principal de uma matriz diferença tal que minimize c(R). Uma breve revisão da literatura sobre a formalização e construção desses critérios é dada a seguir.

1.1 Critérios: RJC, QIC, CIC e $c(\mathbf{R})$

1.1.1 Rotnizky and Jewell's Criterion (RJC)

Rotnitzky e Jewell (1990) propuseram um teste estatístico para a hipótese de que o vetor dos coeficientes de regressão eram iguais a β, e que, se ambos os modelos marginais e matriz de covariância para GEE fossem especificamente corretos, pode-se esperar que Ψ_0 e Ψ_1 são razoavelmente idênticos, em que se define respectivamente como segue:

$$\Psi_0 = K^{-1}\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i \quad , \tag{3}$$

$$\Psi_1 = K^{-1}\sum_{i=1}^{K}\mathbf{D}_i^T\mathbf{V}_i^{-1}(\mathbf{Y}_i-\mu_i)(\mathbf{Y}_i-\mu_i)^T\mathbf{V}_i^{-1}\mathbf{D}_i \quad , \tag{4}$$

$$\Psi = \Psi_0^{-1}\Psi_1. \tag{5}$$

Quando a estrutura de correlação de trabalho é corretamente especificada, Ψ deverá está próxima da matriz identidade. Hin, Carey e Wang (2007) descreveu o critério Rotnizky and Jewell's Criterion(RJ) para selecionar a estrutura de correlação de trabalho como:

$$RJ(\mathbf{R}) = \left[(1-tr(\Psi)/p)^2+(1-tr(\Psi^2)/p)^2\right]^{\frac{1}{2}}, \tag{6}$$

em que p é o número de covariáveis envolvidas no modelo.

1.1.2 Quasi-likelihood under the independence model criterion (QIC)

Para seleção de modelos o critério AIC é bastante conhecido. Contudo, ele não pode ser utilizado para abordagem GEE, visto que é baseado em verossimilhança. Assim, Pan (2001),

propôs um critério com base em quase-verossimilhança para auxiliar na escolha do melhor modelo ou estrutura de correlação cuja expressão é dada por:

$$QIC(\mathbf{R}) = -2\mathbf{Q}(\hat{\beta};\mathbf{I},\mathbf{D}) + 2tr(\hat{\Omega}\hat{\mathbf{V}}_G(\mathbf{R})), \tag{7}$$

em que $\hat{\mathbf{V}}_G(\mathbf{R})$ representa a matrix de covariância estimada a partir da estrutura de correlação de trabalho assumida, $\Omega = \sum_{i=1}^{K}(\mathbf{D}_i^T\mathbf{A}_i^{-1}\mathbf{D}_i|\mathbf{R})$ e, se a matriz de trabalho utilizada é a independente, $\mathbf{R} = \mathbf{I}$, sendo os pares de observação $(\mathbf{Y}_{it}, \mathbf{X}_{it})$ em $\mathbf{D}$ independentes, então a quase-verossimilhança com base em $\mathbf{D}$ é:

$$\mathbf{Q}(\beta,\phi;\mathbf{I},\mathbf{D}) = \sum_{i=1}^{K}\sum_{t=1}^{n_i} Q(\beta,\phi,(\mathbf{Y}_{it},\mathbf{X}_{it})) \tag{8}$$

e assim, define-se o critério de seleção em Hardin (2005) em que $\Omega = \sum_{i=1}^{K}(\mathbf{D}_i^T\mathbf{A}_i^{-1}\mathbf{D}_i|\mathbf{I})$. Utilizaremos o critério de seleção QIC proposto por Pan (2001).

1.1.3 Critério de Informação de Correlação (CIC)

O critério CIC usa metade do segundo termo do QIC para a seleção da estrutura de correlação de trabalho no GEE:

$$CIC = tr(\hat{\Omega}\hat{\mathbf{V}}_G(\mathbf{R})) \tag{9}$$

O primeiro termo do critério QIC, que se baseia em quase-verossimilhança, está livre tanto da estrutura de correlação de trabalho como da verdadeira matriz de covariância. Desta forma, não fornece informação sobre a seleção da estrutura de covariância. Por outro lado, o segundo termo no QIC contém informações sobre a estrutura de correlação através do estimador de variância de sandwich. Embora o segundo termo desempenhe um papel como uma penalização para a seleção de variáveis de modelo marginal, o QIC é mais "pesado" devido primeiro termo. Dessa forma, o QIC não é uma medida particularmente sensível para seleção da estrutura de correlação de trabalho (HIN; WANG, 2009).

1.1.4 Gosho's criterion (c(R))

Como critério de seleção para a matriz de correlação, Gosho, Hamada e Yoshimura (2011) propuseram escolher entre as estruturas de matrizes avaliadas àquela que minimize $c(\mathbf{R})$, representada na equação:

$$c(\mathbf{R}) = tr\left[\left\{\left(\frac{1}{K}\sum_{i=1}^{K}(\mathbf{Y}_i-\mu_i)(\mathbf{Y}_i-\mu_i)^T\right)\left(\frac{1}{K}\sum_{i=1}^{K}\mathbf{V}_i\right)^{-1}-\mathbf{I}\right\}^2\right], \tag{10}$$

em que *tr* refere-se a soma dos elementos da digonal da matriz e $\mathbf{I}$ é a matriz identidade.

Convém ressaltar que em nenhum dos critérios mencionados são incorporados as estimativas limitantes da matriz de covariância, sendo portanto fortemente influenciados pela magnitude das estimativas dos parâmetros de regressão. Dada essa deficiência, a contribuição deste trabalho é pautada na proposta de um novo critério de seleção para matriz de correlação de trabalho, com base nas estimativas limitantes, $\alpha_0(\rho)$, que motivaram a obtenção dos parâmetros β. Sua performance foi avaliada em dois cenários de simulações sob abordagem GEE para respostas correlacionadas normais e binárias.

2 Critério $JCC(\mathbf{R})$: novo critério de seleção da matriz de correlação de trabalho com a incorporação de $\alpha_0(\rho)$

2.1 Motivação e definição

Nos estudos da eficiência sob abordagem de equações de estimação generalizadas, Sutradhar e Das (2000) reportaram a avaliação computacional das matrizes de correlações para as estruturas permutável e AR(1) utilizando a matriz de correlação $\mathbf{R}(\alpha_0(\rho))$, ao invés da matriz de correlação estimada $\mathbf{R}(\hat{\alpha};\hat{\beta})$. Em relação aos aspectos da eficiência do estimadores dos parâmetros da regressão β, dado que a matriz de correlação permutável seja a verdadeira, $\hat{\alpha}(\beta)$ converge para $\alpha_0(\rho)$, satisfazendo a equação $\alpha_0(1-\alpha_0)^{-1}\{t-(1-\alpha_0^t)/(1-\alpha_0)\}-t(t-1)\rho/2=0$, em que $-1/(t-1)\leq\rho\leq 1$, para a suposição de que a matriz de correlação de trabalho seja AR(1). E caso a estrutura correta seja $AR(1)$ e se suponha ter matriz de correlação de trabalho permutável, o estimador $\hat{\alpha}(\beta)$ converge para $\alpha_0(\rho)$, satisfazendo $\alpha_0=2\rho\{t-(1-\rho^t)/(1-\rho)\}/t(t-1)(1-\rho)$ tal que $-1\leq\rho\leq 1$. Dessa forma, passando a fazer uso de $\mathbf{R}(\alpha_0(\rho))$ ao invés de $\mathbf{R}(\hat{\alpha};\hat{\beta})$ para uma dada matriz de correlação $\mathbf{C}(\rho)$.

Seguindo essas especificações, a proposta do critério se baseia na eficiência dos parâmetros de regressão sob avaliação computacional das matrizes de covariâncias utilizando $\mathbf{R}(\alpha_0(\rho))$, no que diz respeito às estruturas de correlações, independente com a matriz identidade, permutável com $\rho_{itt'}=\alpha_0(\rho)$, e AR(1) com $\rho_{itt'}=\alpha_0(\rho)^{|t-t'|}$.

No que segue, propõe-se selecionar a estrutura de correlação que minimiza $JCC(R)$ como matriz de correlação de trabalho representada por:

$$\mathbf{V}_0 = \frac{1}{K}\sum_{i=1}^{K} \hat{\mathbf{V}}_i^{-1}(\hat{\beta})(\mathbf{Y}_i - \mu_i(\hat{\beta}))(\mathbf{Y}_i - \mu_i(\hat{\beta}))^T \hat{\mathbf{V}}_i^{-1}(\hat{\beta}) \quad , \tag{11}$$

$$\tilde{\mathbf{V}} = \frac{1}{K}\sum_{i=1}^{K} \hat{\mathbf{A}}_i^{-1/2}\mathbf{R}_i^{-1}(\alpha_0(\rho))\hat{\mathbf{A}}_i^{-1/2} \tag{12}$$

$$JCC(\mathbf{R}) = tr\left[\left(\mathbf{V}_0^{-1}\tilde{\mathbf{V}} - I\right)^T\left(\mathbf{V}_0^{-1}\tilde{\mathbf{V}} - I\right)\right], \tag{13}$$

em que I se refere a matriz identidade e tr é a soma dos elementos da diagonal da matriz.

Note que quando o $JCC(\mathbf{R}) = 0$, indica que $\mathbf{V}_0^{-1}\tilde{\mathbf{V}} = I$ e consequentemente, $\mathbf{V}_0^{-1} = \tilde{\mathbf{V}}$. Com isso, $JCC(\mathbf{R})$ fornecerá uma medida da qualidade das estimativas dos parâmetros $\alpha(\hat{\beta})$, computados na matriz de correlação para obtenção das estimativas da matriz de covariância para cada indivíduo, $\hat{\mathbf{V}}_i$. Dado a forma intratável de expressar analiticamente a inversa da matriz $\mathbf{V}_0^{-1}$ e o produto $\mathbf{V}_0^{-1}\tilde{\mathbf{V}}$, as propriedades assintóticas podem ser facilmente observadas computacionalmente, de modo que para valores de $-1/2 < \alpha_0(\rho) < -1/3$ ficam impossibilitadas a obtenção das estimativas $\hat{\alpha}(\hat{\beta})$ (CROWDER, 1995).

2.2 Propriedades teóricas das matrizes $\mathbf{V}_0$ e $\tilde{\mathbf{V}}$

As propriedades assintóticas das matrizes $\mathbf{V}_0$ e $\tilde{\mathbf{V}}$ que compõem a formalização do critério $JCC(\mathbf{R})$ são investigadas nessa seção. Analisou-se a convergência de $\mathbf{V}_0$ com base nas estimativas limitantes $\alpha_0(\rho)$. Desde que $\hat{\alpha}(\beta)$ converge para $\alpha_0(\rho)$, segundo as condições definidas em Sutradhar e Das (2000) e citadas na anteriormente, define-se $\bar{\mathbf{R}}_i(\rho) = \mathbf{C}_i(\rho)$, $i = 1,2,\ldots,K$, como a matriz de correlação verdadeira e o estimador $\hat{\mathbf{R}}_i(\hat{\alpha};\hat{\beta})$ de $\bar{\mathbf{R}}_i(\rho)$ pode ser expresso por:

$$\hat{\mathbf{R}}_i(\hat{\alpha};\hat{\beta}) = \hat{\mathbf{A}}_i(\hat{\beta})^{-1/2}\varepsilon_i(\hat{\beta})\varepsilon_i(\hat{\beta})^T\hat{\mathbf{A}}_i(\hat{\beta})^{-1/2}, \tag{14}$$

tal que $\varepsilon_i(\hat{\beta}) = \mathbf{Y}_i - \mu_i(\hat{\beta})$ e sejam

$$\hat{\mathbf{R}} = \frac{1}{K}\sum_{i=1}^{K}\hat{\mathbf{R}}_i(\hat{\alpha};\hat{\beta}) \quad e \quad \bar{\bar{\mathbf{R}}} = \frac{1}{K}\sum_{i=1}^{K}\bar{\mathbf{R}}_i(\rho) \tag{15}$$

de modo que para os resíduos normalizados, $y_i^* = \hat{\mathbf{A}}_i(\hat{\beta})^{-1/2}\varepsilon_i(\hat{\beta})$, $E(y_i^* y_i^{*T}) = \bar{\mathbf{R}}_i(\rho)$. E sob as condições definidas em Balan, Schiopu-Kratina et al. (2005), a saber:

$(C1)$: existe $\delta \in (0,2]$ tal que $sup_{i\geq 1} E(||y_i^*||^{2+\delta}) < \infty$, em que $||y_i^*|| = \lambda_{max}\{y_i^* y_i^{*T}\}^{1/2}$, em que λ_{max} é o maior autovalor;

$(C2)$: $\frac{1}{K}\sum_{i=1}^{K} \mathscr{V}_i \xrightarrow{P} 0$, em que para observações independentes $\mathscr{V}_i = E(y_i^* y_i^{*T}) - \bar{\mathbf{R}}_i(\rho)$,

segue que $\hat{\mathbf{R}}$ converge em média para $\bar{\bar{\mathbf{R}}}$, ou seja

$$\hat{\mathbf{R}} - \bar{\bar{\mathbf{R}}} \xrightarrow{L^1} 0 \quad \text{(elemento a elemento).} \tag{16}$$

Dessa forma, dada as condições citadas, sejam $\hat{g}_i(\hat{\beta}) = \hat{\mathbf{V}}_i{}^{-1}(\hat{\beta})\varepsilon_i(\hat{\beta})\varepsilon_i(\hat{\beta})^T\hat{\mathbf{V}}_i{}^{-1}(\hat{\beta})$ tal que

$$\hat{g}(\hat{\beta}) = \mathbf{V}_0 = \frac{1}{K}\sum_{i=1}^{K} \hat{g}_i(\hat{\beta}) \tag{17}$$

e dada a convergência em (16), tem-se que $E(\hat{g}_i(\hat{\beta})) = \hat{\mathbf{A}}_i(\hat{\beta})^{-1/2}\bar{\mathbf{R}}_i(\rho)\hat{\mathbf{A}}_i(\hat{\beta})^{-1/2} = \tilde{\mathbf{V}}_i$, consequentemente, $\hat{g}(\hat{\beta})$ converge em média para $\frac{1}{K}\sum_{i=1}^{K} \tilde{\mathbf{V}}_i$, elemento a elemento, desde que $\hat{\alpha}(\beta) \longrightarrow \alpha_0(\rho)$.

3 VALIDAÇÃO DO CRITÉRIO JCC(R) POR SIMULAÇÃO MONTE CARLO

Para validação do critério proposto, procedeu-se com estudos de simulações para avaliar a performance do critério $JCC(\mathbf{R})$ em comparação com os $CIC(\mathbf{R}), QIC(\mathbf{R}), RJC(\mathbf{R})$ e $c(\mathbf{R})$. Com esse propósito, realizou-se dois cenários: resposta normais e binomias, computando a taxa de acerto, que indica a proporção em que o critério evidenciou a seleção da verdadeira estrutura de correlação.

Para respostas normais, considerou-se distribuição normal multivariada com média μ_{it}, variância σ_{it}^2, e coeficiente de correlação $\rho_{itt'}$. A média marginal $\mu_{it} = \beta_0(t-1) + \beta_1 x_{it}$ com $\beta_0 = 1$ e $\beta_1 = 0,2$, variância $\sigma_{it}^2 = 1$, e variáveis binárias x_{it}. Para as respostas binomiais, distribuição binomial multivariada com média μ_{it}. O modelo marginal considerado foi $logit(\mu_{it}) = \beta_0 + \beta_1(t-1) + \beta_2 x_{it}$, com $\beta_0 = 0,25$ e $\beta_1 = \beta_2 = -0,25$, sob suposição de verdadeira estrutura de correlação permutável ou AR(1).

Em ambas as situações, os cenários de avaliações foram definidos pela combinação dos fatores tamanhos amostrais, $K = 20$ e 50, o número de observações(t), fixado em $t = 6$, para cada grupo de K indivíduos e, as correlações ρ fixadas em $0,1, 0,3$ e $0,5$.

A obtenção das estimativas limitantes da matriz de covariância, $\alpha_0(\rho)$, foi dada por um processo iterativo que considerou a estrutura de correlação de trabalho permutável e AR(1), satisfazendo as respectivas equações, $\alpha_0(1-\alpha_0)^{-1}\{t-(1-\alpha_0^t)/(1-\alpha_0)\}-t(t-1)\rho/2=0$, em que $-1/(t-1) \leq \rho \leq 1$, sob a suposição de matriz de correlação permutável ser a verdadeira e, supondo verdadeira a matriz de correlação de trabalho AR(1), $\alpha_0 = 2\rho\{t-(1-\rho^t)/(1-\rho)\}/t(t-1)(1-\rho)$ tal que $-1 \leq \rho \leq 1$, conforme descrito na motivação da proposta do critério $JCC(R)$. Dessa forma, as estimativas $\alpha_0(\rho)$ configuram resultados na Tabela 1.

Tabela 1 – Estimativas de $\alpha_0(\rho)$ para $t = 6$

Valores de ρ	Estimativas de $\alpha_0(\rho)$ para verdadeira estrutura de correlação	
	Permutável (EX)	AR(1)
0.1	0,0362	0,2427
0.3	0,1306	0,5360
0.5	0,2688	0,7188

4 RESULTADOS E DISCUSSÕES

Em concordância com os objetivos propostos, para todos os cenários avaliados, o critério $JCC(\mathbf{R})$ apresentou taxas de acerto bem superiores, em relação as taxas observadas no critério $QIC(\mathbf{R})$. Em relação aos demais critérios, as taxas foram similares, considerando pequenas oscilações devido ao efeito do tamanho amostral e grau de correlação ρ.

No tocante ao comportamento dos critérios quando o modelo GEE é ajustado para respostas normais e binomiais, os resultados encontram-se descritos respectivamene nas Tabelas 2 e 3. Na Tabela 2 para pequenas amostras, $K = 20$, a proporção de seleção para estrutura de correlação AR(1) é maior para o critério proposto do que os demais critérios. O critério $QIC(\mathbf{R})$ resultou em um pior desempenho, com taxas percentuais estimadas em $2,6\% - 5,1\%$.

Tabela 2 – Proporções (%) de seleção para estrutura de correlação para respostas normais

Estrutura verdadeira	ρ	Critérios	Tamanhos amostrais (K)					
			$K = 20$			$K = 50$		
			IN	EX	AR	IN	EX	AR
EX	0,1	$JCC(R)$	20,9	47,3	31,8	30,6	42,4	27,0
		$CIC(R)$	40,6	39,6	19,8	45,3	39,9	14,8
		$RJC(R)$	33,0	31,7	35,3	28,5	44,7	26,8
		$QIC(R)$	94,7	5,10	0,20	98,9	1,10	0,0
		$C(R)$	47,8	20,7	31,5	64,2	24,1	11,7
	0,3	$JCC(R)$	18,0	57,9	24,1	13,2	75,3	11,5
		$CIC(R)$	31,2	62,7	06,1	22,8	74,3	2,90
		$RJC(R)$	25,1	39,3	35,6	15,7	63,9	20,4
		$QIC(R)$	97,0	3,00	0,0	99,4	0,50	0,10
		$C(R)$	35,6	51,3	13,1	39,6	47,6	12,8
	0,5	$JCC(R)$	7,70	78,4	13,9	6,00	90,2	3,80
		$CIC(R)$	12,5	85,7	1,80	3,3	96,7	0,00
		$RJC(R)$	25,3	45,9	28,8	15,3	69,0	15,7
		$QIC(R)$	97,4	2,40	0,20	99,6	0,30	0,10
		$C(R)$	29,1	62,1	8,80	18,1	75,8	6,10
$AR(1)$	0,1	$JCC(R)$	8,30	34,3	57,4	3,40	17,2	79,4
		$CIC(R)$	24,1	33,5	42,4	20,4	21,4	58,2
		$RJC(R)$	33,9	31,2	34,9	16,8	35,5	47,7
		$QIC(R)$	90,0	9,10	0,90	98,4	1,40	0,20
		$C(R)$	29,9	19,1	51,0	30,4	12,2	57,4
	0,3	$JCC(R)$	0,30	35,2	64,5	0,00	18,4	81,6
		$CIC(R)$	8,60	47,9	43,5	2,30	43,0	54,7
		$RJC(R)$	20,8	35,0	44,2	4,80	34,4	60,8
		$QIC(R)$	96,0	3,20	0,80	98,7	0,80	0,50
		$C(R)$	8,30	41,2	50,5	1,90	43,1	55,0
	0,5	$JCC(R)$	0,30	47,8	51,9	0,00	37,8	62,2
		$CIC(R)$	3,30	77,6	19,1	0,90	83,5	15,6
		$RJC(R)$	23,0	37,7	39,3	5,90	41,4	52,7
		$QIC(R)$	96,2	3,00	0,80	98,1	1,70	00,2
		$C(R)$	3,40	66,5	30,1	0,20	71,8	28,0

IN: independente; EX: exchangeable(permutável); AR(1)

Verificou-se também que para valores menores de ρ, de modo geral, esse critério ao ser utilizado na seleção da verdadeira estrutura de correlação, os resultados evidenciaram uma tendência em indicar a escolha da estrutura independente. Esses resultados confirmam as observações realizadas por Hin e Wang (2009) de que o critério $QIC(\mathbf{R})$ para pequenos valores de ρ não é recomendado para discriminar estrutura permutável e AR(1).

Com ênfase em destacar a eficiência do critério proposto $JCC(\mathbf{R})$ em relação ao efeito do tamanho amostral, ressalta-se que as taxas de acerto estimadas quando a estrutura de correlação permutável para $K = 20$, foram verificadas em um intervalo de $47,3\% - 78,4\%$, e $K = 50$, as taxas resultaram em estimativas $42,4\% - 90,2\%$. Ao passo que para estrutura $AR(1)$ os respectivos percentuais foram $51\% - 64,5\%$ para pequenas amostras e $62,2\% - 81,6\%$ para $K = 50$.

Para todas as correlações, o critério $JCC(\mathbf{R})$ apresentou um desempenho superior ao critério $C(\mathbf{R})$, entretanto, em relação aos demais critérios, o critério proposto $JCC(\mathbf{R})$ mostrou-se mais efetivo em discriminar a estrutura permutável da estrutura independente, bem como apresentou resultados semelhantes aos demais critérios para as amostras de tamanho $K = 50$, demonstrando sua eficiência assintótica.

Em se tratando do modelo GEE para respostas binomiais, o desempenho do critério $JCC(\mathbf{R})$ comparado aos demais critérios, inicia-se com a discussão dos resultados descritos na Tabela 3.

Os resultados observados na Tabela 3, evidenciaram que os percentuais de acertos dos critérios $CIC(\mathbf{R})$, $RJC(\mathbf{R})$ e $QIC(\mathbf{R})$ quando se consideram $\rho = 0,1$ e $K = 50$, indicavam que a melhor escolha para matriz de correlação de trabalho seriam as matrizes de estruturas independentes, evidenciando assim, o conservadorismo na má especificação da estrutura de correlação independente, quando arbitrariamente supõe-se valores de ρ pequenos.

Para valores de $\rho = 0,5$, as maiores proporções em destaque foram para o critério $C(\mathbf{R})$ na seleção para estrutura AR(1), confirmando os resultados observados em Gosho, Hamada e Yoshimura (2011) em que tal critério para elevadas correlações (ρ) e grandes amostras foi recomendado como melhor para a identificação da estrutura AR(1) em relação aos demais critérios. Igualmente ao que ocorre para o $JCC(R)$, quando se considera valores de ρ pequenos para pequenas e grandes amostras.

Os critérios em geral, fazem uso das propriedades assintóticas dos estimadores $\hat{\alpha}(\hat{\beta})$ de modo que o estimador sandwich (LIANG; ZEGER, 1986), é unicamente determinado pelas estimativas $\hat{\beta}$. Com isso, a matriz de covariância estimada, $\hat{\mathbf{V}}_i(\hat{\alpha}(\hat{\beta}))$, é obtida a partir de $\hat{\alpha}(\hat{\beta})$ e não do valor que motivou sua estimação, o limitante $\alpha_0(\rho)$. Diante disso, a perda da eficiência dos estimadores de β pode ocorrer não somente pela má especificação da estrutura de correlação, como também pelos múltiplos valores de $\hat{\alpha}(\hat{\beta})$ para um mesmo ρ (SUTRADHAR; DAS, 2000).

Tabela 3 – Proporções(%) de seleção para estrutura de correlação para respostas binomiais

Estrutura verdadeira	ρ	Critérios	Tamanhos amostrais (K)					
			$K=20$			$K=50$		
			IN	EX	AR	IN	EX	AR
EX	0,1	$JCC(R)$	11,2	51,7	37,1	23,6	69,0	7,40
		$CIC(R)$	37,9	41,2	21,7	67,6	28,0	5,90
		$RJC(R)$	46,9	24,5	28,6	66,5	25,1	25,1
		$QIC(R)$	61,3	34,7	4,00	90,7	9,30	0,10
		$C(R)$	20,3	59,2	20,5	38,3	55,4	6,30
	0,3	$JCC(R)$	4,50	68,2	27,3	3,80	91,4	4,80
		$CIC(R)$	24,7	55,2	20,2	45,4	50,2	4,60
		$RJC(R)$	43,4	28,3	28,3	49,9	38,0	12,1
		$QIC(R)$	73,0	25,8	1,20	91,8	8,10	0,10
		$C(R)$	3,70	72.1	24,2	6,00	92,0	2,00
	0,5	$JCC(R)$	7,50	55,7	36,8	1,60	97,2	1,20
		$CIC(R)$	17,6	48,2	34,2	29,5	59,6	11,0
		$RJC(R)$	45,5	27,1	27,4	40,1	37,6	22,3
		$QIC(R)$	75,4	22,8	1,80	90,2	9,8	0,00
		$C(R)$	0,20	35,8	64,0	0,20	82,4	17,4
$AR(1)$	0,1	$JCC(R)$	14,1	36,8	49,1	17,1	70,9	12,0
		$CIC(R)$	46,2	24,3	30,6	58,7	28,9	15,1
		$RJC(R)$	43,6	29,8	26,6	56,2	36,3	7,50
		$QIC(R)$	71,5	23,4	5,10	87,2	12,4	0,40
		$C(R)$	29,1	39,7	31,2	32,6	58,8	8,60
	0,3	$JCC(R)$	8,40	37,2	54,4	2,70	48,5	48,8
		$CIC(R)$	29,2	26,7	44,3	29,1	27,7	43,8
		$RJC(R)$	41,0	33,7	25,3	23,0	46,0	31,0
		$QIC(R)$	77,9	19,5	2,60	89,4	10,1	0,50
		$C(R)$	15,9	48,8	35,3	5,60	71,0	23,4
	0,5	$JCC(R)$	2,60	21,1	76,3	1,50	42,8	55,7
		$CIC(R)$	17,3	15,5	67,3	6,30	7,40	86,5
		$RJC(R)$	30,3	29,8	39,9	9,30	26,7	64,0
		$QIC(R)$	78,5	18,1	3,40	90,4	8,70	0,90
		$C(R)$	4,30	22,7	73,0	0,30	9,70	90,0

IN: independente; EX: exchangeable(permutável); AR(1)

Os critérios $CIC(\mathbf{R})$, $RJC(\mathbf{R})$, $C(\mathbf{R})$, e $QIC(\mathbf{R})$ utilizam as estimativas dos parâmetros $\alpha(\beta)$ como medida de seleção da verdadeira estrutura de correlação. Dessa forma, tais critérios ficam unicamente determinados pelas estimativas de β, estimadas pela escolha arbitrária do parâmetro de correlação. No critério JCC, propõe-se a verificação da composição da matriz que identificará a escolha da estrutura de correlação segundo as estimativas do parâmetro β e $\alpha_0(\rho)$.

Convém ressaltar que o critério $JCC(\mathbf{R})$, igualmente ao $C(\mathbf{R})$, não se destina a ser usado para seleção de covariáveis, diferentemente do critério *QIC* que poderá ser usado para escolha do melhor conjunto de covariáveis para um modelo GEE, bem como para seleção da estrutura de correlação de trabalho. Porém, em estudos de investigação sob a performance dos critérios citados, Gosho, Hamada e Yoshimura (2011), apresentaram resultados que confirmam os percentuais apresentados pelo critério $QIC(\mathbf{R})$ e quando comparado ao critério $CIC(\mathbf{R})$, $QIC(\mathbf{R})$ apresenta o pior desempenho na identificação da verdadeira estrutura de correlação, apontando resultados similares em Hin e Wang (2009) no que diz respeito a dependência do critério sob a magnitude dos parâmetros β.

5 APLICAÇÕES

5.1 Aplicação 1

Para exame da aplicabilidade do critério proposto neste artigo, $JCC(\mathbf{R})$, utilizou-se um subconjunto de dados reportado em Hardin (2003) de um estudo longitudinal dos efeitos da poluição do ar sobre a saúde de crianças. A variável resposta são variáveis indicadoras para presença ou ausência de ruído ao respirar, medida em quatro anos cujas idades das crianças são 9, 10, 11 e 12. A covariável fumante identifica o status do tabagismo materno no primeiro ano do estudo.

O modelo marginal com as covariáveis é dado por:

$$logit[E(Y_{it})] = \beta_0 + \beta_1 \text{Cidade}_i + \beta_2 \text{Idade}_{it} + \beta_3 \text{Fumante}_{it} + \beta_4 \text{IF}_{it}, \tag{18}$$

em que Y_{it} são respostas binárias para ausência ou presença de ruído ao respirar para a $i-$ésima criança no tempo t; $Cidade_i = 0,1$ representa a cidade da criança residente em Portage ou Kingston; $Idade_{it} = 9,10,11$ e 12; $Fumante_{it} = 0,1$ representa o status de fumante da mãe da $i-$ésima criança e; IF_{it} representa a medida do hábito de fumante da mãe da $i-$ésima criança no $t-$ésimo tempo de observação.

Ajustou-se o modelo utilizando três estruturas para matriz de correlação de trabalho - independente, permutável e AR(1) e, para avaliar a performance do critério proposto sob a especificação de $\alpha_0(\rho)$, considerou-se os valores de $\alpha_0(\rho) = 0,16082$ e $\alpha_0(\rho) = 0,3544$.

As estimativas dos parâmetros da regressão, erro padrão robusto, estimativas dos parâmetros de associação $\hat{\alpha}(\hat{\beta})$, e valores de $JCC(\mathbf{R})$, $CIC(\mathbf{R})$, $RJC(\mathbf{R})$, $QIC(\mathbf{R})$ e $C(\mathbf{R})$ foram obtidos usando cada uma das matrizes de correlação de trabalho, sendo descritas na Tabela 4.

De acordo com a Tabela 4, as estimativas para idade usando as estrutura independentes e AR(1) são similares, porém para independente, β_2 apresenta menor erro padrão. Ao passo que para as estimativas, β_3 e β_4 as quais encontram-se informações sobre as variações entre crianças para ausência ou presença de ruído ao respirar, e variações para ausência ou presença de ruído da $i-$ésima criança medida sob o hábito de fumante da mãe , apresentam concentração de menores erros padrões quando se faz uso da estrutura de correlação AR(1).

Tabela 4 – Estimativas dos parâmetros β, estimativas $\hat{\alpha}(\hat{\beta})$ e valores dos critérios para três matrizes de correlação de trabalho para ausência ou presença de ruído ao respirar

Covariáveis	Especificação da estrutura de correlação		
	Independente (Erro padrão robusto)	Permutável (Erro padrão robusto)	Ar (1) (Erro padrão robusto)
Intercepto	-0,60123 (0,9460)	-0,55366 (0,9265)	-0,89862 (0,8903)
Cidade	0,14334 (0,6998)	0,08826 (0,7013)	0,36314 (0,6745)
Idade	-0,15995 (0,4065)	-0,21262 (0,4072)	-0,16859 (0,4230)
Fumante (entre indivíduos)	-0,05178 (0,9017)	-0,06994 (0,8981)	0,24713 (0,8445)
Idade:Fumante (dentro indivíduo)	-0,05056 (0,5512)	0,00750 (0,5712)	-0,06401 (0,5641)
$\hat{\alpha}$	0	0,1497	0,3258
	Valores de $\rho = 0$	Valores de $\rho = 0,16082$	Valores de $\rho = 0,3544$
$JCC(R)$	4,2563	2,1334	1,1889
$CIC(R)$	0,45922	0,46217	0,4431
$RJC(R)$	1,4581	2,2476	3,1703
$QIC(R)$	14,0791	14,0889	14,1736
$C(R)$	1,0730	0,86179	0,4077

Os erros padrões robustos para todos os efeitos foram um pouco diferentes. Em particular, os maiores erros padrões para as variações entre e dentro do indivíduo foram observadas sob suposição de correlação independente e simétrica composta, respectivamente.

Os valores de $JCC(\mathbf{R})$ indicam pela seleção da estrutura de correlação AR(1), semelhante ao que ocorre para $CIC(\mathbf{R})$ e $C(\mathbf{R})$, mas não para $QIC(\mathbf{R})$ e $RJC(\mathbf{R})$. Tais resultados confirmam as análises pela escolha da estrutura $AR(1)$ como sendo a mais apropriada.

5.2 Aplicação 2

Para um segundo exame da aplicabilidade do critério JCC, utilizou-se um conjunto de dados em análise sensorial de cafés especiais obtidos da realização do projeto "Protocolo de identidade, qualidade e rastreabilidade para embasamento da indicação geográfica dos cafés da Mantiqueira" aprovado no edital CNPq/MAPA 064/2007 (BOREM, 2007).

O experimento em análise sensorial de cafés especiais foi realizado ao longo de quatro safras (2010/11, 2011/12, 2012/13 e 2013/14), em lavouras comerciais de propriedades localizadas no município de Carmo de Minas, Minas Gerais, Brasil.

O conjunto de dados é formado por quatro variedades de cafés especiais, Bourbon amarelo, Catuaí amarelo, Acaiá vermelho e Mundo Novo. Foram realizadas 288 degustações para cada uma das safras, oriundas de duas altitudes(inferior e superior a 1.200m) em duas formas distintas de processamento (Via seca e úmida) e formadas por dois grupos de vertentes(Sol e sombra).

O modelo marginal para as notas dadas aos cafés especiais pelos provadores na $i-$ésima degustação, $i = 1,2,\ldots,288$, avaliadas nos tempos $t = 1,2,3,4$, com as covariáveis altitudes, vertentes, processamento e genótipo é dado por:

$$\mu_{it} = \beta_0 + \beta_1 \text{Altitude}_{it} + \beta_2 \text{Processamento}_{it} + \beta_3 \text{Vertente}_{it} + \beta_4 \text{Genótipo}_{it} \qquad (19)$$

As estimativas dos parâmetros da regressão, erro padrão robusto, estimativas dos parâmetros de associação, $\hat{\alpha}(\hat{\beta})$, e os valores de $JCC(R)$, $CIC(R)$, $RJC(R)$, $QIC(R)$ e $C(R)$ foram obtidos usando cada uma das três matrizes de correlação de trabalho estudadas nesse trabalho, sendo apresentadas na Tabela 5.

Note que os erros padrões foram diferentes, e que o ajuste utilizando a estrutura independente apresentou menores erros padrões das estimativas. Somente os valores do critério $C(R)$ indicaram que a melhor estrutura para a matriz de correlação de trabalho seria AR (1), enquanto que o critério RJC(R) indicou que a melhor estrutura seria a permutável.

Tabela 5 – Estimativas dos parâmetros β, estimativas $\hat{\alpha}(\hat{\beta})$ e os valores dos critérios para as três matrizes de correlação de trabalho para as notas dadas aos cafés especiais em um experimento de análise sensorial

Covariáveis	Especificação da estrutura de correlação		
	Independente (Erro padrão robusto)	Permutável (Erro padrão robusto)	$Ar(1)$ (Erro padrão robusto)
Intercepto	81,1637 (0,8845)	82,3720 (0,9644)	81,1895 (0,8913)
Altitude			
$> 1.200m$	2,0653 (0,3640)	2,0891 (0,3980)	2,1003 (0,3646)
Processamento			
Cereja descascada	-0,5058	-0,2928	-0,4732
Via úmida	(0,4175)	(0,4818)	(0,4204)
Vertente			
Sombra	-0,0838 (0,4155)	-0,0521 (0,4507)	-0,0737 (0,4176)
Genótipo			
Catuaí amarelo	1,7151 (0,3600)	0,8431 (0,3153)	1,6508 (0,3580)
$\hat{\alpha}$	0	0,2159	0,2248
	$\alpha_0 = 0$	$\alpha_0 = 0,1754$	$\alpha_0 = 0,2237$
$JCC(R)$	9,7459	11,3435	10,6221
$CIC(R)$	3,3758	3,7572	3,3914
$RJC(R)$	1,3938	1,3454	1,3822
$QIC(R)$	17519,14	17882,72	17522,17
$C(R)$	9001,5	9177,221	5841,917

Os valores dos critérios JCC(R), $CIC(R)$ e $QIC(R)$ apontam para a escolha da estrutura independente, confirmando a análise de que as associações das degustações realizadas entre uma safra e outra são independentes.

6 CONCLUSÃO

A performance do critério proposto para dados normais na identificação da estrutura de correlação AR(1) teve destaque para valores de ρ pequenos se comparado aos demais critérios. Apresentou desempenho satisfatório para grandes amostras com valores de ρ maiores, ao passo que para pequenas amostras, destacou-se por diferenciar as estruturas independente e permutável para valores de ρ pequenos. O critério apresentou os maiores percentuais de seleção da

verdadeira matriz de correlação na identificação das estruturas simétrica composta e AR(1), para respostas binomiais, para valores crescentes de ρ.

A magnitude do critério não depende de β, e a robustez do critério proposto pode ser verificada em estudos de simulação.

Agradecimentos

Os autores agradecem pela disponibilidade dos dados referentes ao Projeto protocolo de identidade, qualidade e rastreabilidade para embasamento da indicação geográfica dos cafés da mantiqueira, aprovado no edital CNPq/MAPA 064/2007.

7 REFERÊNCIAS

ALBERT, P. S.; MCSHANE, L. M. A generalized estimating equations approach for spatially correlated binary data: Applications to the analysis of neuroimaging data. **Biometrics**, [Wiley, International Biometric Society], v. 51, n. 2, p. 627–638, 1995. ISSN 0006341X, 15410420. Disponível em: <http://www.jstor.org/stable/2532950>.

BALAN, R. M.; SCHIOPU-KRATINA, I. et al. Asymptotic results with generalized estimating equations for longitudinal data. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 33, n. 2, p. 522–541, 2005.

BOREM, F. M. **Projeto protocolo de identidade, qualidade e rastreabilidade para embasamento da indicacao geografica dos cafes da mantiqueira**. [S.l.], 2007.

CAREY, V.; ZEGER, S. L.; DIGGLE, P. Modelling multivariate binary data with alternating logistic regressions. **Biometrika**, Biometrika Trust, v. 80, n. 3, p. 517–526, 1993.

CROWDER, M. On the use of a working correlation matrix in using generalised linear models for repeated measures. **Biometrika**, Biometrika Trust, v. 82, n. 2, p. 407–410, 1995.

FITZMAURICE, G. M.; LAIRD, N. M. A likelihood-based method for analysing longitudinal binary responses. **Biometrika**, JSTOR, p. 141–151, 1993.

GOSHO, M.; HAMADA, C.; YOSHIMURA, I. Criterion for the selection of a working correlation structure in the generalized estimating equation approach for longitudinal balanced data. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 40, n. 21, p. 3839–3856, 2011.

HARDIN, J. **Generalized estimating equations. Hardin JW, Hilbe J. Generalized estimating equations**. [S.l.]: New York: Chapman & Hall, 2003.

HARDIN, J. W. **Generalized estimating equations (GEE)**. [S.l.]: Wiley Online Library, 2005.

HIN, L.-Y.; CAREY, V. J.; WANG, Y.-G. Criteria for working correlation structure selection in gee. **The American Statistician**, v. 61, n. 4, p. 360–364, 2007.

HIN, L.-Y.; WANG, Y.-G. Working correlation structure identification in generalized estimating equations. **Statistics in medicine**, Wiley Online Library, v. 28, n. 4, p. 642–658, 2009.

LIANG, K.-Y.; ZEGER, S. L. Longitudinal data analysis using generalized linear models. **Biometrika**, Biometrika Trust, v. 73, n. 1, p. 13–22, 1986.

PAN, W. Akaike's information criterion in generalized estimating equations. **Biometrics**, Wiley Online Library, v. 57, n. 1, p. 120–125, 2001.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2015. Disponível em: <https://www.R-project.org/>.

ROTNITZKY, A.; JEWELL, N. P. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. **Biometrika**, Biometrika Trust, v. 77, n. 3, p. 485–497, 1990.

SUTRADHAR, B. C.; DAS, K. Miscellanea. on the efficiency of regression estimators in generalised linear models for longitudinal data. **Biometrika**, Biometrika Trust, v. 86, n. 2, p. 459–465, 1999.

SUTRADHAR, B. C.; DAS, K. On the accuracy of efficiency of estimating equation approach. **Biometrics**, Wiley Online Library, v. 56, n. 2, p. 622–625, 2000.

WANG, Y.-G.; CAREY, V. Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. **Biometrika**, Biometrika Trust, v. 90, n. 1, p. 29–41, 2003.

WANG, Y.-G.; LIN, X. Effects of variance-function misspecification in analysis of longitudinal data. **Biometrics**, Wiley Online Library, v. 61, n. 2, p. 413–421, 2005.

ZHAO, L. P.; PRENTICE, R. L.; SELF, S. G. Multivariate mean parameter estimation by using a partly exponential model. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 805–811, 1992.

CONSIDERAÇÕES FINAIS

O presente estudo apresentou uma nova aborgadem na análise de experimentos provenientes de uma análise sensorial, cujas respostas são pontuadas em uma escala entre zero e dez pontos, e mostrou uma nova perspectiva de estudos das associações entre as respostas categóricas ordinais ao longo de quatro safras. Para tanto, foi introduzida a modelagem Kappa para medir tais associações. Em particular, o primeiro artigo oferece aspectos metodológicos que poderão ser aplicados em estudos que envolvam a avaliação sensorial, a outros produtos, cujas peculiaridades encontradas no conjunto de dados avaliados neste trabalho, também estejam presentes.

Ainda no primeiro artigo, foi possível verificar que, mesmo com genótipos ausentes em uma das safras, o ajuste do modelo marginal para obtenção das probabilidades das notas dadas aos cafés especiais, são perfeitamentes viáveis.

Contudo, a estratégia apresentada no primeiro artigo mostrou-se eficaz para identificação dos atributos qualitativos, que são semelhantes entre as safras, de modo que é possível identificar as diferenças entre os genótipos avaliados pela medida de concordância entre provadores.

A estratégia de modelagem para dados ordinais provenientes de uma análise sensorial, presente no primeiro artigo, consiste na aplicação da metodogia GEE, utilizando três equações de estimação. A metodologia GEE, faz uso da matriz de correlação de trabalho. Portanto, no segundo artigo abordou-se a importância da seleção da referida matriz, e apresentou-se o critério *JCC* para a escolha da matriz de correlação de trabalho, com base nas estimativas limitantes dos parâmetros de associação.

Na proposta do segundo artigo, foi possível identificar que o critério *JCC* é competitivo em relação aos demais critérios apresentados, e que para grandes amostras mostrou-se possuir melhor desempenho. Quando a estrutura da matriz de correlação de trabalho permutável, foi considerada como a estrutura verdadeira nas amostras simuladas, apresentou maiores percentuais.

Dessa forma, vale ressaltar que não consta na literatura estudos em análise sensorial aplicada aos cafés especiais, semelhante ao desenvolvido nesta tese. Logo, a importância deste trabalho aos estudos de análise sensorial para os cafés especiais se dá em duas vertentes: a primeira, por inserir no âmbito de análise sensorial a metodogia GEE para dados ordinais, bem como fornecer ferramentas que auxiliem na identicação das covariáveis qualitativas que

possivelmente afetam o sabor dos cafés especiais, uma vez que o conceito final dado às amostras é proveniente de um conjunto de aspectos qualitativos, associando as relações entre degustações e safras.

A segunda vertente, ocorre pelo fato de que as associações entre as degustações avaliadas ao longo do tempo, com a metodologia GEE utilizando a medida kappa em uma terceira equação de estimação, inseridas neste trabalho foram contempladas em análises estatísticas, possibilitando novos trabalhos no sentido de resolver questões teóricas sobre a captação das estimativas do coeficiente de correlação de concordância. E por que não, construir uma metodologia para identificação do padrão da mudança dos conceitos(notas) dados aos cafés especiais avaliados para períodos entre safras, considerando outras covariáveis além das sensoriais, por exemplo altitude e processamento.

No âmbito da contribuição desta tese na área da estatística, está no fato de que há propostas de trabalhos futuros relacionados a estimação dos parâmetros de associações em duas direções. E no que se refere ao critério *JCC*, a contribuição está sob os aspectos da abordagem de que, se as estimativas dos parâmetros de associação convergem para alguma estimativa limitante, então tal estimativa deve contribuir para escolha da matriz de correlação de trabalho. Essa abordagem, não foi aplicada aos demais critérios, pelo fato de que eles não foram construídos para esse fim. E para concluir, há proposta de estudos relacionados às modificações dos critérios da literatura para uso da abordagem de estimativas limitantes.